STATISTICAL METHODS IN RESEARCH

STATISTICAL METHODS IN RESEARCH

by

Palmer O. Johnson, Ph.D.
Professor of Education
University of Minnesota

PRENTICE-HALL, INC.
New York

First Printing...........May, 1949
Second Printing......January, 1950

PRINTED IN THE UNITED STATES OF AMERICA

To Hilde

PREFACE

If the instruction of students is to keep pace with the rapid development of statistical science, frequent publication of books based on the most recent knowledge in the field is required. Therefore, the author's primary aim is to supply students with a book that is built on the recent advances in statistical theory and practice. Because they approach the study of statistics with different interests, aims, and backgrounds, it is not feasible to write one text that can meet the requirements of all classes of students. Whatever their approach, these students will have one thing in common, namely, they will need to acquire a thorough functional understanding of statistical principles to make intelligent use of statistics. The differences in interpretation of authors of statistical texts as to what functional understanding involves seem to range between the belief that statistical principles are working rules to be learned as quickly as possible for their utilitarian value and the conviction that an advanced knowledge of pure mathematics is the first requisite for the exposition of statistical principles.

The author does not believe that either of these points of view is best for most of the students who need statistical training for their work. The former is likely to lead to blind, rule-of-thumb application of statistical formulas; the latter is indispensable only for those who are to become professional statisticians or mathematical statisticians. Neither practical applications nor mathematical analysis is excluded from this book. In fact, problems have been used abundantly to illustrate principles or results. Also, a number of problems have been inserted whereby the student may test his understanding of the statistical theory. The author is convinced that the detailed working through of problems is fundamental to a functional understanding of statistical techniques. Similarly, application of the principles underlying the design of experimental or observational projects is necessary if a thorough grasp of these principles is to be secured. Experiences in application are necessary for the student if he is to design effective experiments of his own or to evaluate those of others. However, the problems are considered as auxiliaries to the study of the principles.

Again, the mathematical analysis is not excluded because without mathematics there could be no serious study of statistics. But mathematics has been viewed as the servant and not as the master. The question of how much knowledge of mathematics ought to be assumed is difficult to answer. Not many students in the social and biological sciences have a knowledge of calculus. It is the view of the author that the.student should have at least a background in calculus to be able to follow the theoretical material, which cannot be advantageously treated without considerable use of calculus. The understanding of statistical as well as other scientific principles is relative, dependent upon what intelligence, technical background, and experience the student may have. Students with more knowledge of mathematics usually gain more complete understanding of the mathematical formulation underlying a particular statistical principle. They should, therefore, have the opportunity of utilizing their more adequate preparation. However, the number of students with special mathematical training is very limited. But the student with, for example, no calculus, may omit the few sections of the book in which the calculus is used. Even without these sections, he should be able to acquire a considerable and continuous knowledge of the essentials of statistics from the non-technical, logical treatment accorded to most of the essentials in the book.

It should also be explained where the book starts and where it ends. This book does not start from the very beginning of its subject. Many upper-class students and most graduate students have had an introduction to statistics, usually called descriptive statistics, dealing with the elementary processes in the reduction of data. Such preliminary training is assumed. If the student does not have it, the instructor may prefer to begin the subject by laying this elementary base himself. This book deals with the principal objective of statistics, which is to provide indispensable tools and methods for designing and executing experimental and other observational projects and for analyzing and interpreting the results.

It would be advantageous to allot a full academic year to the objectives of statistical methods as presented here. However, adjustment can be made when less time is available by selecting certain portions regarded as most fundamental by the instructor.

It was decided to bring the book to a close when its purpose was accomplished, that is, after the common principles of statistics had been investigated. The aim was not to present topics of interest to a few students only.

The book is based on the content of a year's course in statistics, and was developed over a period of approximately ten years, primarily for graduate students in education and in psychology. During this time, content and method were continuously revised in light of experiences and scientific development of the subject.

The author considers himself especially fortunate in having been a volunteer worker at the Galton Laboratory for a year, during which time he studied with Professor R. A. Fisher, foremost in laying the foundations of modern statistical methods. During this period he also profited from the lectures of and conferences with Professors J. Neyman and Egon S. Pearson.

The chief sources of information and help in developing the book have been the many serious students whose criticisms and reactions were of inestimable value in attaining a clearer presentation of statistical methods. Most significant was the help from my very capable assistants. Among these should be especially mentioned Dr. Cyril Hoyt, Dr. Fei Tsao, Dr. Garland Kyle, and Mr. Stanley Clark, all of whom have made direct contributions to this work.

I am greatly indebted to Dr. Robert W. B. Jackson of the University of Toronto for his critical reading and constructive criticisms of the work in manuscript from which I received valuable suggestions for its improvement.

I am especially grateful to the following authors and publishers for their kind permission to reproduce certain tables which are given in the Appendix:

(1) I am indebted to Professor R. A. Fisher and Dr. F. Yates, also to Messrs. Oliver and Boyd Ltd., Edinburgh, for permission to reprint Tables No. III, Distribution of t, and Table No. IV, Distribution of χ^2 from their book, *Statistical Tables for Biological, Medical, and Agricultural Research;*

(2) Professor George W. Snedecor and The Iowa State College Press for permission to reproduce Table 10.7—5% and 1% Points for the Distribution of F from *Statistical Methods* (Fourth Edition), 1946;

(3) Professor Egon Pearson, Editor of *Statistical Research Memoirs* to reproduce Table IV—5% limits for L_1 and Table V—1% limits for L_1 computed by P. P. N. Nayer.

Palmer O. Johnson

TABLE OF CONTENTS

STATISTICAL METHODS IN RESEARCH

CHAPTER I

THE REALM OF STATISTICS

Statistics in Daily Lif

Our entrance into and departure from this world are recorded as statistical events. Birth and death, marriage and divorce, the school attendance of our children, the crops grown by farmers, the number of miles flown by commercial planes, the hours of our labor, the output of manufacturing plants, the acres of wood demanded for paper, the hours of sunshine, the inches of snowfall—all such events and activities are recorded somehow and somewhere. Myriads of such experiences and events affecting the daily lives of roundly two billion human beings lie behind the statistical data condensed in volumes, published and unpublished. In reverse, we are daily translating into their real meaning statistical data obtained from newspapers, radio reports, lectures, books, and conversations. We act in accordance with the reality implied in statistical data when we conserve fuel which is going to be scarce, when we ship wheat which will be necessary after a poor harvest in a foreign country, when we take precautionary measures against a disease of which unusually many cases have been recorded.

The conception of statistics as having to do with figures is the most popular one, and for good reasons. The public is constantly exposed to statistical data occurring in advertisements, in arguments, and in the distribution of information. If something is said to have been statistically proved, opposition is supposed to become quiescent. Everywhere the ordinary citizen needs some ability to distinguish between what is truth and what is falsehood. In a democracy he needs it most where he participates in the settlement of public problems and contributes toward the growth of public opinion. Citizens not only should be able to look at controversial questions scientifically and dispassionately; they should also acquire the habit of doing so. Education should prepare them to cope intelligently with the problems of their lives and times; they must learn not only to think for themselves but likewise to act for themselves. There is danger in the educational system of a democracy when materials and methods of instruction are not keyed to the formation of the scientific attitude and to the development of the ability to use the scientific method. The ability to use and scrutinize data, to look beneath the surface of things and to discern relations between reality and given data, affords an important safeguard against the dangers of omnipresent propaganda. The problem is to educate man so that he would rather be guided by fact than by emotion.

There is a noticeable similarity between arithmetic and statistics with respect to use in daily life. Arithmetic is so woven into the fabric of our daily life and thought that we use it very often and almost subconsciously. With respect to statistics we need only to recall such phrases as "highly exceptional," "relatively constant," "increases the probability," and "on the average." However, arithmetic is a subject taught in all elementary schools, whereas statistics is taught scarcely at all, although its content is likely no more difficult than that of arithmetic at the same educational level.

The practice of applying certain statistical methods, however simple they may be, is a critical social need for all. We must not forget that even the specialist lives in general society during at least two-thirds of the time. For this longer period he is a layman and needs the best possible quality of layman's understanding. For his guidance in the current of general human living, he needs statistical training.

The most important and undisputed use of statistics in daily life is connected with all the activities of political, social, and commercial institutions which determine the economic and cultural life of a nation. In the realm of policy it is the function of statistics to measure the importance of various problems and to place them in a proper perspective. In many branches of government factual data already are governing policy to a great extent. For instance, the decision to build a number of new schools and to engage more teachers implies legislative measures which are based on statistical investigations of the school-leaving age, the rising birth rate, the increase of population through migration, and other factors. Problems in the economic, industrial, and social fields, such as increase or decrease of employment, shortage of houses, expansion or contraction of existing plants, decrease or increase of crime—these and thousands of others should be solved statistically before political action can be considered. The whole structure of the national budget depends on the sound appraisal of the relationship between potential sources of revenue and planned expenditure. Local authorities need statistical information for the districts they serve; national agencies need it for the country; the organization of the United Nations needs it for the world.

It is essential that governmental agencies be prepared to make the fullest possible use of modern statistical methods. The public is entitled to the benefit that may be derived from the progress in research. Old methods are often wasteful or have been found unreliable. One should expect that government, the foremost user of statistics on a large scale, should pioneer in the application of modern statistical methods.

The urge to apply modern statistical developments seems to be greater where an immediate personal advantage is involved in commercial life.

There is even one branch of commercial activity which owes its existence and all-pervasive development to statistics: insurance. In many other branches a combination of technical and statistical knowledge is used. The planning of a large factory or combine is now a part of what is known as "scientific management." Many firms have planning departments which use statistical returns and charts to a great degree. An unusual example is furnished by the seemingly sentimental enterprise of manufacturing greeting cards, of which approximately three billion are mailed each year in the United States, involving an annual sale of 135 million dollars and postal charges of 100 million dollars. "Statistical planning," taking place in a special department one and a half years ahead of the exhibition of a card in a store is the first step toward the sale. Everywhere knowledge and experience are needed for planning production, distribution, and sales, although the statistical methods used are often not very elaborate. In administration, statistics provide measures of performance and efficiency. Although the data do not state the causes of inefficiency, if any, and do not directly effect improvement, they are pointers; their value depends entirely on the use which is made of them.

Underlying all planning is the guidance derived from statistical data of the past toward the goals desired for the future. An insurance company quoting rates for an endowment life policy to mature twenty years hence can and must do so on the basis of an estimate of future interest rates and past mortality experiences. The size of a new factory is determined partly by estimates of future demands for the products to be manufactured. Most goods for consumption are made or ordered long before they are sold. Consumers, nowadays starting their own organizations, no less than producers and managers, are dependent, for forceful action, on the instruments provided by statistical methods. Thus it is profitable to be able to forecast trends for all economic groups: for business management comprising large firms with international, long-range distributions as well as for the individual merchant supplying the immediate needs of a local neighborhood.

On the other side, employees everywhere are finding that it is of vital importance to labor and its aggregate organizations to use statistics, which represent tools in the formation of their organizations and programs.

The United States excels in using methods for forecasting trends in every field of industry and public life.

Statistics in the Sciences and the Arts

Statistical devices have made their greatest advances in the scientific and technical branches of industry, where enterprise and science not only meet but are amalgamated.

Perhaps no branch of mathematical science has had a more rapid growth than has the science of statistics. In the span of the last sixty or eighty years the methods of statistics and the probability calculus have infiltrated one branch of science after another, until they now hold a central position in physics, biology, meteorology, chemistry, and astronomy. Furthermore, statistics is also growing in significance in a number of other fields, such as the political and social sciences. To what may this remarkable growth most likely be attributed?

The introduction of a new theoretical device into a field of knowledge may often seem incidental in that when it first becomes available, it is used when it appears to be of value, just as the microscope, X rays, or integral equations may be tried out. In the case of statistics, however, its introduction was not just casual.

At first statistics was used apologetically, perhaps with the excuse that it was only an expedient to help overcome a temporary shortcoming, as in reducing large amounts of observational material in order to study details. Thus at first the new "weapon" was tried with the expectation that it could be used in the study of detail, as in the study of hereditary transmission in individuals from one generation to another, or, as in physics, to fill in the gap in knowledge in gas theory with respect to initial coordinates and velocities of the single atoms.

Attitudes in scientific research shift at times, perhaps unintentionally. Interest in individuals shifted to the mechanism underlying the behavior of aggregates of individuals. It was suddenly realized that even if the individual case could be studied in detail, it would be necessary to follow up thousands of individual cases in order ultimately to integrate them all into one statistical enumeration.

Charles Darwin was fully appreciative of the essential function of statistics in biological study. His theory depended on the law of large numbers. Every living species is continually producing a multitude of individuals. On the whole, the better-fitted ones live more abundantly and have a better chance of survival. The large geometrical progression of potential offspring and the enormous destruction of actual offspring to be inferred from it constitute the statistical mechanism operating to produce the very small increase in the chances of survival that a small favorable variation bestows.

The change in the status of statistics as a subordinate device was most drastic in physics. Here it came to take the dominating role of defining the goals and showing the ways of reaching them. Thus the entire structure of science was shaken, since it rests upon the foundation of physics. This role of statistics has led to a new understanding of the essential qualities of the laws of nature, namely, the change from a deterministic formulation of laws underlying the occurrence of natural events to one in terms of statistical regularities, based—as in Darwin's theory—on the law of large numbers. This transition from the interpretation of physical laws based on the notion of causality to one derived from statistical theories is attributable largely to Boltzmann's[1] interpretation of the classical law of entropy, or the second law of thermodynamics, as it is usually called. According to this interpretation,. the second law rests upon statistics. Rather, it is statistics; that is, it is a purely statistical law. Heat flows in the direction from higher to lower temperature because the chance is only one in many billions that it is likely to do otherwise. Events go in the direction in which it is most probable that they will move (Ref. 5).

Further developments, particularly the new quantum mechanics and Heisenberg's uncertainty principle, have revolutionized still more the usual conception of the older classical physics and contributed to the building of the edifice of the statistical conception of nature. While these changes have been taking place, the physicists have developed their own statistical methods, particularly quantum statistics, quite apart from the methods of statistics in other fields. Statistical ideas are utilized in some modern chemical theories, such as the structural formula of certain organic substances like rubber and proteins, where chains of molecules of different weights and lengths are postulated. For example, chemical changes in such substances are interpreted as alterations in the frequency distribution of chain length.

The significance of the general philosophical implication of the statistical formulations relating to the construction of scientific theories can hardly be overrated. We are more directly concerned here, however, with pointing out briefly the position that statistics holds today in certain fields of science and in technology. Since about 1920, the statistical approach has been accepted and welcomed by a steadily increasing circle of scientific workers, until today this approach is probably one of the most characteristic features of modern science.

The role of statistics in science begins with the interpretation of measurements. Even though the methods of the natural sciences are the most reliable thus far designed for finding out matters of fact, the conclusions drawn from them are only probable, since they are based on evidence formally incomplete. This fact is statistically described by the attachment of a *coefficient of error* to the measurement.

Take, for example, the measurement of the distance of the sun from the earth, or, speaking more correctly, the semimajor axis of the earth's orbit. This is the most important constant in astronomy, since it establishes the scale not only of the solar system but also of the whole universe. It is used in almost any calculation of distances and masses, of sizes and densities of planets, of their satellites, and of the stars. Therefore, any error in its calculation is multiplied and repeated in many different forms. Its importance has stimulated measurements of ever-

[1] It should be noted that the work of Willard Gibbs followed parallel lines.

increasing accuracy. At present the measurement is 93,005,000 ± 9000 miles (p.e.);[2] that is, the distance is uncertain to 1 part in 10,000. One hundred years ago, the uncertainty was 1 part in 20. The progress in the development of any science is indirectly given by the size of the errors in its measurements.

Laboratory measurements in physics and chemistry are subject to experimental errors. Considerable attention has been given recently to methods of controlling and evaluating all variables that might conceivably influence the results. The purpose is to obtain reliable laboratory standards, such as those of capacity, frequency, and voltage. Particularly with the development of the sciences of biochemistry and biophysics, measurements are required on material essentially variable. A wide field is under development in which are used such statistical methods as sampling, followed by analyzing and testing of the experimental results as well as the closely related problem of appropriate experimental designs. These methods have increasingly important applications in industry. The same situation prevails also to a slight extent in engineering—mostly in technical control and research. Engineers have developed methods of their own for dealing with the variation in the materials which they use. It is likely that the use of statistical methods of treating variations in these fields would be more efficient than the current use of the factors of safety.

Statistical methods are indispensable tools of the industrialist who is concerned with the manufacture or purchase of presumably similar articles or units on a large scale. However efficient the control of production may be, the products are bound to vary, and it is necessary to check the extent of variation by some plan of routine testing. The conformity to the requirements of a consignment of raw or manufactured materials must be reliably established. Considerable headway has been made in recent years in developing efficient statistical methods and experimental designs for meeting requirements. The productive process must be in a state known as one of statistical control, the criterion for which is: the sequence of materials must exhibit the property of randomness. These are statistical problems, for the solution of which the most advanced statistical methods are necessary. At times, when operations were found lacking statistical control, statistical analysis of the results of routine tests have been used successfully to locate the source of the unwanted variations. The application of statistical methods can protect the consumer against the vagaries of sampling and safeguard the producer from the losses incurred by chances "unjust" to him.

Meteorology is a branch of applied physical science which has a statistical basis, since weather forecasting utilizes statistical principles

[2] Probable error.

and methods. The meteorologist collects data which are relatively complex and which are the result of multiple factors operating together without control. Hence, he has to apply methods of multivariate analysis and also statistical methods developed for dealing with serially correlated data. It may be expected that, with the rapid development of electronic calculators, striking improvement will be made in solving the problem of long-range weather forecasting. The great problem in weather forecasting at present is the lack of means to work out all the mathematical variables within the period that knowledge of this kind is useful. If valid predictions could be made of the weather long enough in advance, it might even become possible to do something about the weather. Agriculture, shipping, air travel, and other activities would benefit by advanced knowledge of the weather. The savings in lives, crops, and money would be incalculable.

In practically all branches of biology, methods of statistics are used. Galton, influenced largely by the ideas of Darwin, made quantitative studies of biological variation. Much of the recent development in the theory and application of statistics arose to meet the need for improved tools designed to handle problems in agricultural and biological research. There was a need in these fields not only for interpreting observational data but also for planning experiments efficiently.

Genetics is a branch of biological science which seeks to explain the resemblances and the differences that are displayed among organisms related by descent. Whereas the earlier work in this field was chiefly descriptive and empirical, the development of theories based on Mendel's discoveries has brought statistical methods to bear more and more on the problems. In fact, highly developed statistical methods now constitute the basis of an important part of the subject. The once conflicting sciences of biometry and genetics are now closely integrated.

Public health, epidemiology, and vital records are statistical in character. The collection and analysis of large masses of data are fundamental in those fields. Federal and state governments collect data for informative and directive purposes. The study of population changes is somewhat specialized; its facts are the facts of life on which scientific planning for the future depends. Populations are recruited by birth and depleted by death.[3] The balance between them and the change in character of the age-group patterns of the population are subjects requiring careful and critical statistical analysis. Statistical methods are increasing in use in research in many branches of medicine, though apparently the general practitioner has not been greatly affected by statistical ideas. Statistical methods are also fundamental in the standardization of biological extracts. In biological assays, such as in

[3] Immigration has, of course, been an important factor in the United States.

the calculation of the potency of penicillin, insulin, digitalis, and other drugs, the necessary precision could not be realized except by the use of modern statistical procedures.

The psychologist, particularly in the fields of experimental and applied psychology, needs a working knowledge of statistical methods. In a quantitative inquiry into a psychological problem it is generally necessary to measure a limited number of cases. In selecting them the psychologist must be sure that they are effectively representative of the population from which they are drawn. Usually at least two samples, namely, experimental and control groups, are necessary in an experiment. These must be so selected as to eliminate any bias of selection with respect to characteristics that are related to the investigation. In addition, the problems of measurement involve the determination of the reliability and validity of the instruments used. Finally, analyzing the experimental data and drawing conclusions that the data merit are essentially statistical procedures.

Applied psychology emphasizes the importance of individual differences; it needs to develop tests for intelligence, skills, and aptitudes of various kinds. The allocation of individuals to places in society for which they are best fitted requires tests of mental and physical traits. The statistical methods of multivariate analysis are essential for the interpretation and use of such data. The future of human civilization depends to a great extent on the capacity of man to understand the factors and forces governing or controlling his own behavior. In the solution of these problems statistical method is likely to play a significant role.

Psychologists have developed from orthodox statistical methods some variants of their own. The methods of factor analysis, for example, are used to describe the human mind by means of a small number of psychological factors.

One of the earliest uses of the term *statistics* was the description, at first verbal and later numerical, of outstanding characteristics of a state. The interpretation given in the first issue of the *Journal of the Royal Statistical Society* (Ref. 8) is: "Statistics may be said . . . to be the ascertaining and bringing together of those facts which are calculated to illustrate the condition and prospects of society." Social science was the parent of statistical method. A characteristic of the method of the social scientist was the restriction of his observations to circumstances that were not amenable to experimentation. Hence he usually dealt with complex cases of multiple causation. The science of economics is perhaps the best example of this use of statistics.

Tippett (Ref. 6) gives three reasons why economics is dependent on statistics. One reason is that economic laws, if they exist, pertain to mass or group phenomena. The preferences, desires, and reactions of millions of people are manifested in economic events. The so-called "law of supply and demand" applies very widely. The fundamental assumption underlying the existence of sciences like economics (and psychology) is that statistical laws are descriptive of human behavior. A similar assumption underlies a rational approach to business and political problems. The second reason for the dependence of economic science on statistics is that only quantitative data, that is, statistics, can yield laws in the scientific sense. The third reason lies in the nature of economic problems. Economic experiments are usually not feasible. Hence, if phenomena are to be observed and explained, the method of study is essentially statistical rather than experimental. It is not often possible to isolate one or a few factors for experimental study as is done by the experimentalist in his laboratory.

In economics research there are three general uses of statistics: they may (1) serve as information culminating in hypotheses and theories, (2) be applied to the testing of hypotheses or theories, and (3) furnish estimates of quantities in economic analysis.

There has not been much cooperation between theoretical economists and statisticians. However, the development of statistical methods has been notable in economics. The increasing use of such quantitative concepts as prices, income, and supply and demand may mean that the approach of the statistician eventually will prevail over that of the theorist.

We meet specific problems to which statistical analysis has been applied in telegraph and telephone communication, in electric-power distribution, in road and rail traffic, and so on. The theory of probability has been usefully applied in the study of the effects of chance and other factors in accidents. It has been noted that individuals differ in their proneness to suffer accidents under given conditions. ✓

Statistical facts and methods play a significant part in the development of sociology and education as sciences. The collection of statistics illustrative of the conditions of society has been mentioned as one of the earliest activities. Each national census depicts our industrial, economic, and social status at a given time. Social surveys are frequently conducted in different parts of the country to find out the status of unemployment, housing, the delinquency of youth, and so on. The method of inquiry may be by sample, with its own special difficulties and sources of error. Sociology stresses the interdependence of social facts and the need of considering them in relation to each other. The comparative method used at times applies the principle of varying the circumstances of a phenomenon with a view to eliminating variable and unessential factors. Thus it aims to arrive at what is indispensable and constant. Its primary purpose is to make provision for classification of forms of social relationships to facilitate causal analysis. Statistical investigations of crime, of the causes of suicide, and of the conditions under which certain economic organizations arise illustrate how the comparative method has been applied.

Educational statistics collected by Federal, state, and local authorities provide more and more the basis for educational policies and programs. Subjects illustrative of the amenability of educational problems to scientific study are: changes in the school population with respect to age, intelligence, and other characteristics; means of providing equality of educational opportunities; the location of youth with special talents. Likewise, numerous studies employing the experimental method, particularly those applying modern principles of experimental design, are adding genuine knowledge concerning the educational process.

National opinion polls, such as those of Gallup, Crossley, and *Fortune* magazine, use systematic methods of sampling. The development of this means of measuring public opinion is likely to play a significant part in the theory and practice of democratic government.

Statistics is beginning to find application even in such nonscientific fields as the arts. In a task comparable to that of the telephone engineer who tabulated the frequency of principal words in order to secure the best possible transmission, a literary scholar has tabulated the six thousand most common words in English, French, German, and Spanish. Some points of disputed authorship have been decided by the statistical study of the length of sentences. The frequency with which colors and sound patterns occur in poetry, the number of types of imagery used by Shakespeare, the number of different word classes characteristic of prose and poetry of certain periods—all these are illustrations of statistical applications. Evidence of errors in the chronology of early Roman history has been revealed by certain life tables. The authenticity of paintings has been established by means of the frequency of brush marks.

The work of the mathematical statistician is fundamental in the development of statistical science. Here, as in other fields of science, basic research contributes general knowledge which affords the means of solving a large number of significant practical problems, although a specific solution may not be provided to any one problem. The role of applied research is to discover complete solutions to specific problems. The new knowledge provided by basic research furnishes scientific capital, from which source practical applications must be obtained. Most of the mathematical theory of statistics in its present character is the result of research of recent decades. Perhaps in no field of science have the theoretical advances been so sweeping and the practical results of such advances so pronounced. The reason may be that the solution of theoretical problems was primarily rendered indispensable by the urgent requirements of practical research. Furthermore, the principal contributors to the solution of the theoretical problems discovered the actual need for such solutions in their direct contact with the problems of practical research.

There is usually a gap between theoretical developments and practice in scientific fields, and this gap is also characteristic of statistics. The width of this gap varies in the several applied fields, and there is even a wide variation among workers within the same field with respect to the quality of statistical methods used.

The rapid development of statistical science has, of course, left many problems unsolved, both theoretical and practical. It may be expected that theoretical studies of statistics will increase in the immediate future, leading to greater rigor of its theoretical structure. As is characteristic of all scientific subjects, statistical science is never finished and complete: it is dynamic, developing always. The result will be more and more rigorous methods (Ref. 2). This development is likely also to take in areas and fields where new types of observational data and new kinds of observations will be sought. Also, supplementary mathematical researches will be found necessary before workers in such fields can carry out their studies with the high standards of competence employed in fields where statistical methods are firmly rooted.

Mention should be made of the mechanization of statistical calculations. The generation of calculators as users of logarithms and prepared tables of mathematical functions and other aids has been succeeded by one which knows only how to produce figures mechanically. Commercial machines for accounting and for scientific computation have done much to benefit business, government, and science. It is not only in removing the drudgery of reducing large masses of statistical data that the mechanization of statistics is important: with the development of machines based upon the principles of electronics rather than of the cogwheel, the most complicated and advanced mathematical applications become practically solvable for the first time. The significance of this development for the solution of theoretical as well as practical problems in science is just beginning to be realized. The impetus given to this development by the exigencies of World War II was very great. No matter how rapidly one machine is produced, when finished it seems to be almost obsolete, so swift is progress. Therefore, any description of the electronic calculator which is given here is likely to be soon superseded.

The electronic numerical integrator and computor, the Eniac, invented and perfected at the Moore School of Engineering of the University of Pennsylvania, does not have a single moving mechanical part. Only the tiniest elements of matter—electrons—move within its 18,000 vacuum tubes and several miles of wiring. This amazing machine completes in two hours a mathematical task which 100 trained men could do only in a year. Since all mathematical tasks, however abstruse or involved, can be reduced to basic arithmetic if ample time is provided, this machine practically eliminates time to give the answers to virtually any problem. That is, basically the machine does nothing more than perform the fundamental arithmetic processes. This it does by the generation of very precisely timed electrical impulses. These impulses are formed at a speed of 100,000 per second, which is equivalent to one operation every twentieth impulse, thus adding, for instance, at the rate of 5,000 per second. The Eniac has four kinds of memory. One of these "minds" performs the task of indicating the initial and boundary conditions of the problem. All problems must first be broken down into their essentials, which are then punched on cards. These cards are then run through a machine unit known as the "reader." The reader acts as the translator of the mathematical language to the language of the machine, and vice versa. The values of certain scientific constants are introduced when required. The machine can handle numbers of 20 digits.

Machines have already been planned to solve problems running into 400 stages, that is, machines which have a "memory" of 400 numbers. Such a machine could solve 100,000 different equations in approximately one minute.[4]

The illustrative rather than exhaustive review that has just been presented has attempted to portray the realm of statistics all the way from daily life through theoretical and applied science. If the purpose has been achieved, the all-pervasive character of statistics should be realized. A knowledge of statistics—at least of its logic and its dependence on the data of experience—is indispensable to everyone in the practical affairs of human society. Statistical science has likewise pervaded both the theoretical and applied aspects of the biological, physical, and social sciences. In fact, every observable event in the behavior of man, as well as in the behavior of rocks and stars, is amenable to scientific treatment and correlation with other events. In this analysis, statistical methods have come to play a necessary part if such data are to be assayed with scientific precision and if the reliability of the information is to be determined with objective validity.

The bricks of experience and the mortar of reason are the twin supports upon which the indestructible foundation of science is built. The essence of science is the rational ordering of the facts of experience. In this process the data of experience are represented by concepts. The concepts are defined in a manner which facilitates the interpretation of rational relation between experiences. Although the derivation of these relations involves pure reasoning, statistical methods based on the theory of probability contribute in the drawing of inference and conclusions by specifying the degree of uncertainty involved.

Statistics in all its aspects is accordingly of interest and importance to a large number of classes of people. However, there are few if any individuals, including professional statisticians, who can be experts in

[4] See the discussion of meterology, page 7.

all branches of statistics, because they would then need to be expert in many branches of knowledge, including the foundations of statistical science itself as well as the many fields of application. Statistics is both a science and an art. Statistics is a science because its methods are basically systematic and of wide application. Statistics is an art because success in its application is dependent on the skill, special experience, and knowledge of the person using it in the field to which the application of statistical methods is made. Such qualifications are necessary because the data collected in any field are the manifestations of persons or things with which the statistician needs a first-hand acquaintance.

It is, therefore, of importance that the author of any text in statistical methods make clear the purpose and scope of his book.

STATISTICS IN THIS BOOK

The traditional and popular notion of the function of a statistician is the collecting, tabulating, and describing of long records of figures. These records are conveniently summarized by the calculation of averages, percentages, index numbers, and other descriptive measures, and by the construction of one or more of the kinds of tables, graphs, diagrams, or charts. This process of reducing data to certain summary values has been greatly aided by high-speed machines for tabulation and calculation. The collection of experimental and other observational data is, of course, an indispensable part of the scientist's work. However, the function of a statistician, as now recognized in many branches of science, goes far beyond the collection and processing of numerical data for descriptive purposes. The less widely known activities include his contributions to advances in mathematical statistics basic to the creation of tools of scientific value. These tools give precision to tests of scientific hypothesis. They also indicate how observational studies including experiments must be planned, whether under laboratory, factory, or field conditions, to provide the most reliable and valid information with the least expenditure of time, energy, and money.

The emphasis of this book is on the interpretative rather than on the descriptive function of statistics. This book also aims to present the theoretical foundation of modern statistics, not as an end in itself but principally to provide the background for the intelligent application of modern statistical methods. The medium for developing an understanding of the theoretical foundation is primarily empirical and logical, supplemented at times by mathematical formulation. The complete exposition of the mathematical theory of modern statistical methods is, however, beyond the scope of this volume. Such information would be of interest chiefly to mathematical statisticians, since a thorough understanding of the theory of modern statistical methods requires a fairly advanced knowledge of pure mathematics. Until recently, the basic researches in the mathematical theory of statistics were rather widely dispersed among scientific journals, but books dealing principally with the mathematical theoretical foundation of modern statistics are now available (Refs. 1, 3, 4, and 7). Thus, although from the mathematical standpoint this book is not self-contained, it is written for readers without specialized mathematical training.

The theoretical presentation in this book has been based, as it must be for present-day needs, on original and secondary sources of mathematical statistics. It is assumed that certain aspects of this theoretical background must be clearly understood if statistical methods are to be put to intelligent use. One basic conception is that one must know how to choose the most effective statistical tool for the purpose in mind. A second is that one must know the basic assumptions underlying the statistical tool selected. A third basic conception is that one must first test to see if the assumptions are fulfilled by the particular situation to which the tool is to be applied. By continuous emphasis in this text upon these requisites, it is proposed that the user of statistical methods will become habituated to the practice of critical examination and selection rather than to applying statistical methods blindly or in a rule-of-thumb manner.

Let us repeat: statistical method is based on the same fundamental ideas and processes as is the general scientific method. Thinking statistically is equivalent to thinking scientifically. This kinship underlies the development of the principles of statistical methodology. The more complete understanding of scientific methods is a direct aim in the presentation of this text. Reasoning skepticism, scientific caution, and common sense are urgently needed in statistics.

The more significant contributions to statistics since the early 1920's have been made in the development of the foundations for the problem of statistical inference. The principles of statistical inference deal with two chief problems: that of testing statistical hypotheses and that of statistical estimation. These, then, are the two fundamental statistical problems of the research worker. The presentation of the theoretical aspects of these two problems, with special emphasis on their practical aspects, constitutes the principal content of this book, which has been arranged with the view of presenting the main ideas underlying statistical inference in a logical developmental order leading to a functional understanding of the principles.

The concepts underlying probability and likelihood as they are used in statistics are given first, since probability theory plays the primary role in statistical inference. The fundamental theorems of direct probability follow. We proceed with other theorems which, in turn, lead to the classical binomial, normal, and Poisson distributions.

We then discuss the development of sampling theory and its use in problems of statistical inference. The selection of representative samples receives considerable emphasis in keeping with the requirements of present-day research.

This background should prepare the student to understand the testing of statistical hypotheses. Many illustrations of current procedures of testing statistical hypotheses are presented. The problem of estimating parameters from sample values is then treated. The following are considered: the properties of "best" estimates; the form of the frequency distribution of observational values in relation to the most accurate estimates; and two methods of forming estimates—the method of maximum likelihood and the method of interval estimation. Many original practical problems are worked out by way of illustration.

The interpretative function in statistical analysis has been mentioned as one of major concern. The fact is, however, that the interpretation of a body of data requires a knowledge of how it was obtained. It is of equal importance that conclusions drawn from observational results be based on detailed knowledge of the procedures employed in the investigation. Thus, the major function of a statistician is to design experiments and to plan investigations which will yield maximum information and valid conclusions. This responsibility of a statistician is stressed throughout.

Considerable space has been allotted to the technique of the analysis of variance, the most powerful statistical tool yet devised for analyzing sources of variation. Modern experimental and sampling designs require this technique for the analysis of their results. Related problems such as those in regression are also included.

A thorough understanding of the problems of the field in which one works is essential when statistical data from this field are to be collected and analyzed. To develop statistical craftsmanship, one must acquire skill by observation and much practice. The aim of this book is to assist students and research workers who require technical aid in the design, execution, and interpretation of quantitative researches which may originate in the laboratory or in the field. This book is designed just as much to help a student to become a competent critic of the research literature in his field.

The content of this text is based largely on the theory and application of those statistical methods which are of general importance. The same formula is applicable to diverse groups of subject matter, as is true of other branches of mathematics. The specialized uses of statistics involve no great alteration of structure; rather, the specialization consists in the way in which statistics is applied. No attempt has been made in this book to present illustrations from the many varied fields to which statistical methods can be usefully applied. The student should become competent to deal with many analogous problems through a study of the statistical processes illustrated in the examples. He is, therefore, invited to work through the numerous examples in all numerical detail, so that he may learn how to apply the same methods not only to the unsolved problems given in the text but also to those encountered in his readings and, above all, in his own research.

Much care has been given to the practical arrangements of numerical calculations. The analysis of the results obtained from modern and original experimental designs has been given special attention.

References

1. Cramér, Harald, *Mathematical Methods of Statistics.* Princeton, N. J.: Princeton University Press, 1946.
2. Kendall, Maurice G., "On the Future of Statistics," *Journal of the Royal Statistical Society,* Vol. CV, Part II (1942).
3. ———, *The Advanced Theory of Statistics,* Vol. 1. London: Charles Griffin & Company, Ltd., 1945.
4. *Idem,* Vol. II, 1946.
5. Schrödinger, Erwin, "The Statistical Law of Nature," *Nature,* Vol. 153 (1944), pp. 704–705.
6. Tippett, L. H. C., *Statistics.* London: Oxford University Press, 1943.
7. Wilks, S. S., *Mathematical Statistics.* Princeton, N. J.: Princeton University Press, 1946.
8. Yule, G. Udny, *An Introduction to the Theory of Statistics.* London: Charles Griffin & Company, Ltd., 1929, p. 3.

CHAPTER II

PROBABILITY AND LIKELIHOOD

The work of a scientist is in part practical: he designs experiments and makes observations. Another part of his labor is theoretical: he formulates conclusions from his experimental findings, compares his results with those of other workers, constructs a theoretical system so as to represent and order the facts of observation as accurately as possible, or notes their conformation to existing theory. With the aid of the theory he derives predictions, which he again validates by new observations.

The Basis of Statistical Inference. In most, if not all, of these activities of the modern scientific worker, statistical methods play a significant part. If the experiment or investigation is to lead to explicit, unequivocal, and convincing results, it must be planned so that the data are capable of clear-cut statistical interpretation. The testing of underlying assumptions, the drawing of inferences from sample to population or from observation to hypothesis, and the derivation of predictions are all based upon intelligent statistical analysis.

One of the most hazardous acts of the research worker is the drawing of inferences or conclusions from experimental data. This act is a process of reasoning from the part to the whole, from sample to population, from the particular to the general, or from effect to cause. This step is difficult, it seems, because the experimental results pertain to the experiment or sample, whereas the inference or conclusion refers to the population, of which the experiment or sample is only a very small part.

The inferences drawn from sample to population are uncertain. Even so, these inferences can be rigorous, because they may be made so as to include within themselves a quantitative specification of the kind and amount of uncertainty involved. Upon this achievement depends the validity of the process of acquiring new knowledge by observation or experiment. Science can progress by collecting new experiences as well as by the better ordering of those already possessed. It is primarily by the former process that new knowledge comes into being.

The statistician's contribution to the problem of drawing conclusions from experimental results consists in (a) setting up the requirements for the design or the logical structure of the experiment and (b) interpreting the data. While these two aspects of the process of adding to scientific knowledge are closely related, our principal concern for the present is to consider the general problem of statistical inference. As has been noted in Chapter I, there are two chief problems of statistical inference: that of testing statistical hypotheses and that of estimation. Preliminary to the direct consideration of these problems, it is desirable to develop some fundamental ideas and theorems, which have their origin in probability theory. The interpretation of experimental data is based on the application of probability theory. This theory is planned to provide the mathematical model of the empirical facts, that is, the data with which the statistician works.

Setting up a Model. In looking for a solid theoretical foundation upon which to build a model, the statistician must make clear just how far the concepts which he uses are justified and are requisite. The justification of the logical system he develops rests upon the demonstration of its usefulness in describing the results of experience. The events and objects of the world of reality are always very complex. The scientifically trained mind is required to identify the characteristic or salient point from among the vast number present as an essential condition from the standpoint of theory. Because the objects of the world of reality cannot be comprehended in a way that could lead to an exact theory, they are superseded by idealized conceptions which can be comprehended with comparative ease. The object of creating theoretical models is to permit the mental reconstruction of the world of empirical fact. This statement is not equivalent to saying that the theory necessitates putting the empirical facts into an inflexible predetermined scheme. On the contrary, the theoretical system must be constructed so that the facts are truthfully represented. A scientific theory may be abstract not only in that it encloses a collection of selective facts but also in that it covers a set of ideal objects, such as wave function in physics and the plane in geometry. Yet when such theories encompass real objects to close approximations, they may serve a useful purpose. The statistician begins his work in developing efficient working tools for the research worker by building a simplified model by which he proposes to represent the phenomena of observation with reliability sufficient to supply useful results.

Statistical Interpretation of Probability. The principal function of statistics is to describe certain characteristics of mass phenomena and repetitive events. From the theoretical point of view, unlimited sequences of events or of similar observations are referred to as *statistical universes* or *populations* or *collectives.* Much of theoretical statistics is built up around the idea of an infinitely large hypothetical population of which the observational data make up a sample. The idea of an infinite parent population from which samples are taken is a mathematical abstraction. Populations with which we deal in practice are finite. The infinite population may be considered as a limiting case of a finite population when the number of individuals increases indefinitely. In experimental work, also, a hypothetical infinite population may be considered as an infinite population of all experiments that might have been carried out under the conditions of an observed experiment. The individual experiment is interpreted as a random selection from the infinite population so defined.

A population is an aggregate of individuals. The individual case is of interest to the statistician chiefly because it is from the collection of individuals that the characterization of the population becomes possible. Even if the interest were in the individual case, information would need to be collected for thousands of individuals, and perhaps no other eventual use of them would be made besides combining them under a single statistical generalization. Although in this treatment the identity of any particular individual is irretrievably lost in the aggregate, it does not follow that we cannot say anything about the individual from the knowledge we have of the population. Take, for instance, the frequency distribution of the ages of the 24,395 high-school graduates as recorded in Table 45, page 202. Let us take a single individual from the group or population of 24,395. Even though we do not know his age, we know that he will be an exceptional individual with respect to age if he is of less than, say, sixteen years. It can be said that he will be one of 84/24,395ths of the group. He will, of course, more likely be one of the 12,148/24,395ths of the group. It is more convenient, when dealing with problems of this type, to use a term commonly called *odds* or *probability*. In the illustration just cited, it can be said that the odds or probability of any one individual's being less than sixteen years at the date of graduation from high school is 84/24,395 = .0034, and the probability of his age being eighteen is 12,148/24,395 = .498. This interpretation of probability is the one usually accepted in modern statistics; that is, probability is the ratio of frequencies. As in this illustration, so in any frequency distribution: statistical probability may be considered as the means by which the characteristics of the whole distribution may be ascribed to the random individual.

The long-standing controversy over the nature and meaning of probability need not detain us here. We may merely mention that the psychological and subjective interpretation should be kept distinct from the objective or operational interpretation of relative frequencies. Probability is associated with our subjective sense of expectancy just as a thermometer reading is linked with our subjective sense of heat and cold. The evaluation of probabilities from given data on the basis of standard calculations of secondary from primary probabilities is objective in the sense that this manner of derivation is acceptable to most modern statisticians.

Two definitions of probability may be cited here. (1) Von Mises (Ref. 3) defines the probability of an event as the limit of the relative frequency of this event in an infinite sequence of trials, the *Kollektiv*, fulfilling certain specified conditions. In a purely mathematical sense, the existence of this limit is assumed to be axiomatic. (2) Kolmogoroff (Ref. 2) gives the most comprehensive discussion of probability from the standpoint of measure. He defines probability as a set function which fulfills a certain system of axioms. This theory starts with the concept of the frequency ratio but does not postulate that definite limits of frequency ratios exist. It builds around the concept of a random variable, that is, by considering the probability of an event as a number connected with the event. The axioms postulated in the theory express the principles for operating with the numbers. With respect to application, the two theories are largely equivalent. However, the limiting properties of frequencies involved in definition (1), rather than the pure mathematics of abstract ensembles occurring in definition (2), will be accepted as the basis of the frequency theory of probability insofar as it is used in our present discussion.

Thus, the true probability, P_{12}, of getting a double 6, or sum 12, in one throw of two dice is defined as $\lim\limits_{n \to \infty} \frac{n_{12}}{n}$, assuming that the limit exists, where n_{12} is the number of times a score of 12 is obtained in n throws of the two dice. Similarly, probability values can be determined for each of the other possible totals. On a priori grounds, a tentative or hypothetical probability could be assigned to the true probability. However, probability in the sense used here in statistics depends for its meaning on aggregates of phenomena or repeated events. Although the value of P_{12}, for instance, can never be reached in practice, it can be attained within an arbitrary degree of certainty by making n sufficiently large. According to a theorem by James Bernoulli, the probability that the relative frequency $\frac{n_{12}}{n}$ will be adjacent to P_{12} is arbitrarily near to 1 for a sufficiently long sequence of trials.

EXAMPLE 1. *An Experiment in Probability.* We shall illustrate some of the main points in probability theory by considering an experiment consisting of the throws of a pair of dice. This experiment was repeated a large number of times. The sequence of throws of the pair of dice gives rise to a sequence of numbers, the variable consisting of the sums of the several combinations of the two sets of dots on the two upper faces of the dice after each throw, that is, 2, . . . , 12. The conditions of each throw were kept as uniform as possible. The systematic record of the results of sequences of this kind constitutes a set of statistical data relative to the events observed. Six sets of data, resulting from 36, 360, 3,600, 36,000, 180,000, and 360,000 throws, are recorded in Table 1. The data are arranged in frequency distributions which show the number and per cent of occurrences for each of eleven possible events, 2, . . . , 12.

TABLE 1

DISTRIBUTIONS OF SCORES WITH TWO DICE FOR VARIOUS NUMBERS OF THROWS, COMPARED WITH THEORETICAL EXPECTANCY*

Scores	Theoretical		36 throws		360 throws		3,600 throws		36,000 throws		180,000 throws		360,000 throws	
	f	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%	*f*	%
12	1	2.8	1	2.8	12	3.3	110	3.1	1,044	2.9	5,403	3.0	10,431	2.9
11	2	5.6	2	5.6	13	3.6	214	5.9	2,201	6.1	10,617	5.9	20,503	5.7
10	3	8.3	2	5.6	37	10.3	368	10.2	3,190	8.9	15,296	8.5	30,207	8.4
9	4	11.1	3	8.3	44	12.2	363	10.1	4,118	11.4	20,344	11.3	39,918	11.1
8	5	13.9	4	11.1	57	15.8	515	14.3	5,170	14.4	25,381	14.1	50,352	14.0
7	6	16.7	10	27.8	58	16.1	602	16.7	6,047	16.8	30,059	16.7	60,063	16.7
6	5	13.9	7	19.4	43	11.9	470	13.1	4,885	13.6	24,665	13.7	49,629	13.8
5	4	11.1	3	8.3	47	13.1	408	11.3	3,804	10.6	19,434	10.8	39,919	11.1
4	3	8.3	1	2.8	24	6.7	295	8.2	2,782	7.7	14,401	8.0	29,488	8.2
3	2	5.6	3	8.3	18	5.0	173	4.8	1,896	5.3	9,717	5.4	19,781	5.5
2	1	2.8	0	0.0	7	1.9	82	2.3	863	2.4	4,683	2.6	9,709	2.7
Total	36	100.1	36	100.0	360	99.9	3600	100.0	36,000	100.1	180,000	100.0	360,000	100.1

* The author is indebted for the data to Dean Earl Hudelson, West Virginia University, who performed this experiment.

The hypothetical or theoretical distribution arrived at on a priori grounds is also recorded in the first main column.

The probability of getting a 12 in the throw of two dice shows some fluctuation in the six series of experiments, ranging in value from .028 to .033. A similar situation holds for each of the other totals. The true probability of getting a 12, $\lim\limits_{n \to \infty} \frac{n_{12}}{n}$, though never reached in practice, can be approached closer and closer by increasing the size of n. On this basis, the value .029 determined by 360,000 throws would give the best approximation; likewise for the other totals. In this way probability statements are based on the results of empirical investigations.

The erratic or haphazard behavior of the fluctuations of the variable from throw to throw is usually spoken of as *randomness*. Even with the utmost care in keeping all relevant factors under control, the results vary from observation to observation in such an irregular way that exact prediction of any single event is impossible. The sequence may, therefore, be called a sequence of random experiments. It is noted, however, that, in spite of the unpredictable behavior of individual results, the average results of long sequences of the random experiments exhibit a striking regularity; this regularity may be inferred from the similarities among the several percentage frequency distributions. It is this phenomenon that serves as the basis for the mathematical theory of statistics.

The hypothetical value of probabilities may at times be very useful in furnishing clues to true probabilities.

We may use the theoretical values of P to determine the *mathematical expectation*, a concept that will be encountered later in sampling theory. The mathematical expectation of any quantity is the sum of all the values it may assume multiplied by their respective probabilities:

$$E(X) = P_1X_1 + P_2X_2 + \cdots + P_nX_n = \sum_i P_iX_i \qquad (2.01)$$

Formula (2.01) shows that the mathematical expectation is the weighted arithmetic mean of a variable where the different probability values, P_i's, provide the weights. The mathematical expectation of the throws of two dice is given by:

$$\left.\begin{aligned} E(X) = {} & (\tfrac{1}{36})2 + (\tfrac{2}{36})3 + (\tfrac{3}{36})4 + (\tfrac{4}{36})5 + \\ & (\tfrac{5}{36})6 + (\tfrac{6}{36})7 + (\tfrac{5}{36})8 + (\tfrac{4}{36})9 + \\ & (\tfrac{3}{36})10 + (\tfrac{2}{36})11 + (\tfrac{1}{36})12 = 7 \end{aligned}\right] . \qquad (2.02)$$

Summary. In the interpretation of probability statements on the basis of relative frequencies, the following points are essential (Ref. 4):

(1) The probability of an event has meaning only when the individual event is an element of the specified reference class.

(2) The objective values involved in probability statements grow out of their determination through empirical investigations.

(3) Since probability relates to the property of an object in a specified reference class, a given property can be associated with various degrees of probability referred to different reference classes.

(4) The direct evidence for probability statements is statistical in character, since the definition of such statements is explicitly stated in terms of relative frequencies. There are, however, cases where indirect evidence provides estimates of the probabilities and validation of them: for example, when probability statements are a part of a system of statements.

(5) Every probability statement defined as the limit of a relative frequency is a hypothesis which is incapable of complete confirmation or final verification by means of the finite evidence available at any specified time.

(6) Probability statements to be used successfully for specifying the occurrence of designated properties in definite classes with stable relative frequencies are not dependent upon "deterministic" or "indeterministic" issues.

Fundamental Theorems of Direct Probability. The function of the calculus of probability is to derive probabilities of compound events from sets of initially given probabilities. Thus, in the example of dice casting, given above, the probability of throwing a 6 with a die is not a problem in the calculus of probability; but, given this probability, the probability of getting 12 in the throwing of two dice is such a problem. It should be recognized that the propositions asserted in the calculus are only analytic of the definitions and rules originally specified, as in the case of demonstrative geometry, for instance. The probability calculus thus makes possible the derivation of relative frequencies with which certain events occur from the initial probability statements without the specification in the statements of what the actual frequencies are. In thus making definite the predictions which the probability statements involve, the calculus enables us to make the check of statement content. In this section a few of the standard rules regulating the calculus of direct probabilities will be given. Most of the science of statistics is built upon the explicit or implicit application of these fundamental rules (Refs. 1 and 4).

It is assumed, to begin with, that probability is measurable on a continuous scale. Thus, a probability is a real number, and any two measures of probabilities are comparable, that is, $P_1 > P_2$, $P_1 = P_2$, or $P_1 < P_2$.

The probability of a proposition A on data R is written

$$P\{A|R\}$$

Thus, we may state as

Rule 1. If R entails A, $P(A|R) = 1$
If R entails not—A, $P(A|R) = 0$

Thus, if an event is certain to happen, its probability is 1; if it is certain not to happen, its probability is 0. The range on the probability scale is from 0 to 1. Any value between these limits is, therefore, a positive proper fraction.

Rule 2. If $P_1, P_2, \ldots, P_n$ are the probabilities of n mutually exclusive propositions $A_1, A_2, \ldots, A_n$ on data R, then the probability that one of the propositions is true is $P_1 + P_2 \cdots + P_n$. Symbolically:

$$P\{A_1 \text{ or } A_2, \text{ or } \cdots A_n|R\} = P_1\{A_1|R\} + P_2\{A_2|R\} + \cdots + P_n\{A_n|R\}$$

Thus, if one ball be drawn from a bag containing four white, five black, and seven red balls, since the chance of its being white is $\frac{1}{4}$ and of its being black is $\frac{5}{16}$, the probability of its being either white or black is $\frac{9}{16}$.

Rule 3. The probability of two propositions A and B on data R is the product of the probability of A given R and that of B given A and R. Symbolically,

$$P\{AB|R\} = P(A|R)P(B|AR)$$

More generally,

$$P(A_1A_2 \cdots A_k|R) = P(A_1|R)P(A_2|A_1R)P(A_3|A_1A_2R) \cdots P(A_k|A_{k-1} \cdots A_1R)$$

Thus, the probability of drawing a second white ball from a bag containing five white and four black balls, the ball first drawn being returned before the second drawing, is $\frac{5}{9} \times \frac{5}{9}$, or $\frac{25}{81}$.

The probability of becoming a total orphan is the product of the probabilities of being bereaved of father and of mother.

The rules for the logical sum of events (Rule 2) and for the logical product (Rule 3) are basic in the elementary calculus of probability. From them, by the application of the ordinary rules of logic and arithmetic, it becomes possible to derive significant consequences. One such derivation is Bayes's theorem, which, from the consequences drawn from it, often plays a conspicuous part in treatments of the foundations of probability and scientific method. Symbolically, it may be stated as

$$P\{A_i|RH\} \propto P(A_i|H)P(R|A_iH)$$

That is, the probability of A_i, given R and H, is proportional to the probability of A_i, given H, multiplied by the probability of R, given A_i and H. The factor on the left, that is, $P\{A_i|RH\}$, is called the *posterior* probability; the first factor on the right, $P(A_i|H)$, the *prior* probability; and the remaining factor, $P(R|A_iH)$, the *likelihood*.

In order to make any practical use of Bayes's theorem, it is necessary to decide on the values to be ascribed to the *prior* probabilities. Bayes and Laplace postulated that, in the absence of definite knowledge, the antecedent probabilities were assumed to be equal. This postulate has been relentlessly attacked, especially in recent years, by statisticians on the grounds of supplying by hypothesis data unavailable through empirical or, more particularly, statistical investigations.

The Principle of Maximum Likelihood. Since in most cases it is practically impossible to assign values of empirical significance to the a priori probabilities in Bayes's theorem, the theorem has only a limited use. Therefore, it plays a very minor role as a means for determining the probability of a given hypothesis on the grounds of the available evidence.

Statisticians who reject Bayes's postulate supplant it with a different principle based on the use of likelihood. That is, for any A_i and H,

$$P\{A_i|RH\} \propto P(A_i|H)L(R|A_iH),$$

where the factor $L(R|A_iH)$ stands for the likelihood function.

The principle of maximum likelihcod states that, when the problem of choosing from a number of hypotheses, A_i, arises, we are to choose the one (assuming it exists) that maximizes $L(R|A_iH)$. That is, we are to select the hypothesis which gives the maximum probability of the observed event.

Other Theorems in the Calculus of Probability. The previous rules governing direct probability calculations are based on the assumption that the relative frequency of a proposition referred to a specified class of objects or events has a limit. There are other theorems in the calculus which require the fulfillment of additional assumptions. One of these is that the condition of irregularity obtains in the reference classes. This condition is known as a *random character.* It may be spoken of here as a method of selection which affords an eoual probability to certain propositions and thus permits the application of the calculus probability a priori.

The irregular *Kollektiv,* by which is meant an infinite sequence of observations, is the foundation of the mathematical theory of probability advanced by von Mises. The condition of randomness, or impossibility of a gambling system, which the *Kollektiv* must satisfy, means that if the relative frequency of some particular attribute is calculated in a subsequence of the *Kollektiv,* selected by some method which is independent of the *Kollektiv* itself, it must tend to the same limit as it does in the original *Kollektiv.* Randomness is fundamental in the theory of sampling to be discussed later, since the theory deals principally with samples generated by such processes.

The Binomial Distribution. For reference classes satisfying the condition of random character, the following can be shown: If the probability of having a specified property, say S, called "success" is p, and the probability of not having it is $q = 1 - p$, then the numerical value of the probability that exactly t elements in a set (where $t \leqq n$) have the property S while the remaining $n - t$ elements do not have S is given by

$$P_{n,t} = \frac{n!}{t!(n-t)!}p^tq^{n-t} \left| \begin{array}{l} t = 0, 1, 2, \cdots, n; \\ p = p(S) = \text{constant} \\ \quad \text{for group of trials} \end{array} \right. \tag{2.03}$$

This important theorem is termed the *binomial law.* $P_{n,t}$ is the general term in the binomial expansion of

$$(q + p)^n$$

The maximum value of $P_{n,t}$, where p and n are fixed, varies with t. This maximum value is given when t satisfies the condition

$$pn + p \geq t \geq pn + p - 1 \tag{2.04}$$

When n is very large, the value for which t gives a maximum may be taken as pn. This value indicates that the probability of sets with n successive elements which contain exactly t elements with the property S is largest when t is approximately equal to pn, or that the proportion of S's in a set of n elements is approximately equal to the limit of the relative frequency of S in the *Kollektiv.*

Equation (2.03) is a special case of a more general theorem dealing with situations in which not only two results are considered but in which the event may occur in k ways with probabilities $p_1, p_2, \ldots, p_k$. Then, for a random sample of N from a multinomial distribution, it can be shown that the probability P_k of N giving n_1 of the first kind, n_2 of the second, . . . , n_k of the last, is

$$P_k = \frac{N!}{n_1!, n_2!, \ldots, n_k!}p_1^{n_1}p_2^{n_2} \cdots p_k^{n_k} \tag{2.05}$$

which is the general term in the multinomial expansion of

$$(p_1 + p_2 + \cdots + p_k)^N; \qquad N = n_1 + n_2 + \cdots + n_k$$

The Poisson Distribution. An important distribution of the discontinuous type which often describes the facts of observations is one where p, or the probability of an event, is very small, but where a large number of cases or trials, n, are taken so that pn is finite but small. The number of occurrences will be distributed in the Poisson series. Thus,

$$\begin{array}{ll} p \rightarrow 0 & n \rightarrow \infty \\ q \rightarrow 1 & np \text{ remains finite } = \mu = \text{mean} \end{array}$$

It can be shown that, for the Poisson distribution,

$$\begin{aligned}\text{Mean} &= m\\ \text{Variance} &= npq \rightarrow m\end{aligned}$$

The distribution is therefore determined by one parameter.

If $t = 0, 1, 2, \ldots,$ the relative frequency with which the values occur is given by the series

$$\begin{array}{cccccc} e^{-m}, & \dfrac{me^{-m}}{1!}, & \dfrac{m^2e^{-m}}{2!}, & \ldots, & \dfrac{m^xe^{-m}}{x!} \\ t = 0, & 1, & 2, & \cdots, & x \end{array} \tag{2.06}$$

This series is known as "Poisson's limit to the binomial," "the Poisson series," or "the law of small numbers." Probability tables for the distribution are given by Pearson (Ref. 5).

The Normal Distribution. The binomial law basic in the theory of probability is exact, but it possesses the distinct disadvantage of involving much labor, particularly in the computation of the factorials that enter in the term $P_{n,t}$ [see (2.03)] when n is large. Furthermore, it is a theoretical distribution of the discontinuous or the discrete form. When the character is continuous, as is very often the case in measurements in science, a curve is essential in describing such continuous variation.

It can be shown that by a series of approximations an analytic formula can be obtained from Equation (2.03) which takes the form

$$P_t = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{\delta^2}{2\sigma^2}}, \qquad \begin{aligned}\text{where } \delta &= t - np\\ \sigma &= \sqrt{npq}\end{aligned} \tag{2.07}$$

and the graph of P_t as a function of δ is a symmetrical, bell-shaped curve variously called the normal distribution curve, the Gaussian curve, or the Laplacian-Gaussian error curve. Since the maximum value of the exponential e^{-x}, for $x \geq 0$, is unity, it is noted that the normal approximation for the probability that t will assume its most probable value is given by $\dfrac{1}{\sigma\sqrt{2\pi}}$, or in terms of the binomial parameters, $\dfrac{1}{\sqrt{2\pi npq}}$, where $q = 1 - p$. It is obvious that the normal approximation gives the closest fit to the binomial when $p = q$ (see page 58).

In addition to the normal curve being the limiting form of the binomial distribution, as well as of certain other distributions, its usefulness in theory and practice is especially enhanced by the central limit theorem. According to this theorem, under certain conditions the sum of n independent random variables, in whatever form they may be distributed, tends to be distributed, when expressed in standard measure, as the normal distribution when $n \rightarrow \infty$. Another important property of the normal distribution is its reproductive property. For example, a linear function of variates that are normally distributed is itself normally distributed.

One of the earliest applications of probability was to the systematization of measurements and observations in the physical sciences, particularly astronomy. Legendre in 1806 had formulated what has become known as the principle of least squares: When a set of empirical observations is used to establish the constants of a mathematical function, the best solution is that which reduces the sums of the squares of the residual errors to a minimum. This principle was later placed on a definite mathematical and logical basis by the work of Gauss, Laplace, Maxwell, and others. That is, the normal curve, although previously formulated by de Moivre, was developed as a useful mathematical tool. Since it was used by Gauss to describe the distribution of "errors," it was spoken of as the "normal curve of error." The distribution curve is useful, however, in many situations which have nothing to do with "errors," as in the original setting in which it was used. It is useful in dealing with variations of different kinds, especially with experimental and other observational results, as in the biological sciences.

Serious attempts were made, particularly by Quetelet, to apply the theory of probability to social statistics. He popularized the idea of the "average man" as computed from extensive statistics which he collected. It was through analogy of the average man to the center of gravity in mechanics that he assumed human actions or traits as occurring in accordance with the operation of laws giving rise to a normal distribution. Unfortunately, this attempt, although the influence of Quetelet soon became very slight, seemed to have established the use of the term "normal" in connection with a law of distribution presuming that measurements should always be expected to follow the "normal law of errors" as if it were a law of nature, Though later developments have shown that in science the normal curve gives at times a very close approximation to the observed facts, these instances of very close approximations are the exception rather than the rule.

In dealing with the distributions of errors of measurement or observation, the normal law of error was derived under the assumption that deviations from the most probable value are fortuitious, meaning that the forces in operation to produce them could not be resolved into more elemental factors. It was assumed that the deviations were as likely to be positive as negative, and that they varied without limit, that is, within the bounds of $\pm\infty$. Laplace's generalization was that the distribution obtained by the repetition of a great number of identical alternatives is represented by the function e^{-x}, such that the ordinates of the normal curve decrease on both sides of the maximum ordinate in such a way that their logarithms are proportional to the squares of the distances from the center. Extending this idea to fluctuations other than so-called "error," it may be said that, if an observation, say x, is a resultant of the sum of the effects of a large number of small causes operating at random, and if each effect is independent of x, the obtained distribution is expected to be normal.

The normal distribution holds a central place in the theory of sampling as well as in the theory of probability.

PROBLEMS

Exercises 1–9 are based on the assumption of a normally distributed population.

1. What proportion of the total number of cases lies between one and two standard deviations (S.D.) above the mean?
2. What is the probability of obtaining a value of the variate in random selection at least as large as $+1.96$ S.D.?
3. What proportion of the area under the normal curve lies between 1.27 S.D. and 1.33 S.D.? lies above 1.3 S.D.? lies above -1.3 S.D.? lies below 2.1 S.D.?
4. What is the probability that a measure will lie in the range 2.5 S.D. to 3.1 S.D.?
5. What is the probability of obtaining an absolute value of x/σ greater than 1.5?
6. What is the relative length of the ordinate cutting off the lowest 12.1 per cent of the area?
7. A variate is normally distributed with mean 13.5 and S.D. 3.6. (a) What measures selected at random might be expected to occur in not more than 5 per cent of the cases? in not more than 1 per cent of the cases? (b) What is the probability of obtaining a value of 15? of 8?
8. A variable is normally distributed with unit standard deviation. The probability of obtaining a value of 15 or greater from the population is .132. What is the mathematical expectation of the means of random samples?
9. A population has a mean of 37.6. It is found that 95 per cent of the values of the variate lie in the range 27.8 to 47.4. What values of the variate will occur with a probability of .01 or less?
10. Insofar as the theory of statistics is concerned, upon what does the concept of probability depend for its meaning?
11. In rolling a die, the variable X takes on values 1, 2, 3, 4, 5, and 6. If the die is unbiased, show that $E(X)$ is 3.5, $E(X^2)$ is 15.167, and the standard deviation of X is 1.708.
12. In the classical example of the Poisson distribution given by Bortkiewicz, the records of 20 army corps over a period of 10 years furnish 200 observations of the number of men killed by the kick of a horse. If the number of deaths is denoted by the variable X, which takes the values 0, 1, 2, 3, and 4 with frequencies 109, 65, 22, 3, and 1,

show that the mean is approximately equal to the variance. Find the theoretical frequencies.

13. Calculate the frequency of girls in 100 families of 3 children each; $p = .49$.

14. Find the number of different committees, each of 3 persons, that can be selected from 5 individuals.

References

1. Kendall, Maurice G., The Advanced Theory of Statistics, Vol. I. London: Charles Griffin & Company, Ltd., 1945, pp. 164–185.

2. Kolmogoroff, A., "Grundbegriffe der Wahrscheinlichkeits-rechnung," *Ergebnisse der Mathematik und ihrer Grenzgebiete*, Vol. 2 (1933), No. 3.

3. Mises, Richard von, *Probability, Statistics and Truth.* New York: The Macmillan Company, 1939.

4. Nagel, Ernest, "Principles of the Theory of Probability," *International Encyclopedia of Unified Science*, Vol. 1, No. 6. Chicago: The University of Chicago Press, 1939.

5. Pearson, Karl (ed.), *Tables for Statisticians and Biometricians.* London: *Biometrika*, University College.

CHAPTER III

SAMPLING DISTRIBUTIONS

If the tools designed by the mathematical statistician are to be used intelligently and efficiently by the research worker, the former cannot evade the responsibility of setting forth clearly and unequivocally the conditions under which the use of each tool is valid and efficient. Where the statistician has done his part, it is the responsibility of the research worker to determine whether the necessary conditions obtain in his particular case. It should be pointed out that other tools are generally required to test whether or not these conditions hold good. The command of these tools is an indispensable part of the researcher's art. Once it has been established that the assumptions have been fulfilled, he can proceed with confidence in the results.

So that the student may gain an insight into the logic and reasoning underlying the problems of drawing valid conclusions from experimental results, we present a number of commonly used models developed by the statistician for such purposes. It should be emphasized that the ability to distinguish the specific use or uses for each of the models will go a long way toward developing the kind of statistical craftsmanship essential in the modern research worker.

Preliminary Notions on Sampling and Inference. The material out of which the statistician constructs his model for practical use in interpreting experimental results is discovered by noting what happens when sample after sample is taken from the same population. It is noted, of course, that the results usually differ from one sample to another. Since the method of selection is kept uniform throughout the sampling process, these discrepancies can logically be assigned only to the process, because clearly the population remains constant. It is proper, therefore, to speak of the fluctuations from sample to sample as *sampling errors*. These sampling or *chance* errors, as they are sometimes called, are found to follow chance laws, that is, though all together they form a uniform result, the value any sample might have cannot be accurately predicted. The individual deviations are unanalytic; that is, the forces operating to bring them about are incapable of resolution into simpler and identifiable components. Out of these sampling errors the statistician makes his model. Against such a standard it becomes possible to compare the experimental results. Since it is possible to measure the amount of sampling error to be expected in any given case, it is necessary only to note whether or not the experimental results conform with the standard, that is, to compare the relative magnitude of the experimental results and their random sampling errors. In this comparison, if we note that the observed results, namely, the estimate of an effect presumed to exist, could seldom (say once in a thousand trials, once in a hundred, or once in twenty) be as large or larger owing to random errors of sampling alone, then the effect is said to be *real* in the sense that it is not likely to be due to sampling errors alone, and the experimental results are said to be *significant.* On the other hand, if it is found that often (for instance, fifty times in one hundred, one time in five, or even once in ten, and so forth) results as large or larger could be obtained that would be attributable to random sampling errors alone, they are said to be *insignificant.* Ordinarily, the basis of determining whether results are significant or insignificant is as follows:

(1) The results are said to be significant if the conclusion that they are would be erroneous in 1 per cent or less of the cases.

(2) The results may be significant but further observations are necessary (that is, we suspend judgment) if the conclusion that the results are significant would be wrong in 5 per cent or less but more than 1 per cent of the cases.

(3) The results are not significant if our conclusion that they are significant would be in error in more than 5 per cent of cases.

The technical term for the process employed in examining the significance of experimental or observational results is "the test of significance." This process will be discussed much more completely in Chapter IV, The Testing of Statistical Hypotheses.

The examples of empirical sampling experiments given below illustrate successive stages by which the statistician builds up the statistical models to be used in interpreting experimental results. This method, the way in which the earlier statisticians worked, provides a simple way of understanding quite rigorously the theoretical foundation underlying statistical inference. Today it is not usually necessary to do an actual experiment in order to construct these statistical models based on sampling errors, since the theory of probability enables the statistician to deduce sampling distributions theoretically. In fact, the theoretical deduction of the sampling distributions of the numerous statistical quantities now in use is a highly specialized branch of mathematical statistics. This deduction is sometimes a problem of great mathematical difficulty. Particularly when new types of observational data are under consideration or where information of new kinds is under search, the mathematical problems at times have proved to be so formidable that the statisticians have had to rely on actual sampling experience. Although the mathematical derivations are of fundamental importance to statistical theory and practice, it should be apparent that the conclusions to be drawn from such mathematical models would have no justification beyond the fact that they agree with what actually happens experimentally or would be arrived at from these simple sampling experiments. Theory is a tool which is tested through application and whose usefulness is decided in connection with the application.

TABLE 2

100 RANDOM SAMPLES OF 5 FOR A VARIABLE X FROM A POPULATION WITH MEAN 30 AND STANDARD DEVIATION 10

25	19	26	23	25	50	20	34	5	17	25	29	38	30	34	24	18
28	35	35	29	33	19	32	22	44	18	19	21	23	26	20	30	10
24	25	30	22	30	29	43	36	27	17	35	40	34	19	19	23	28
14	38	22	37	24	33	33	27	47	22	28	26	31	32	30	11	36
24	42	29	47	24	43	16	21	38	35	42	24	47	14	38	28	27
32	26	21	29	36	20	9	46	29	21	33	26	56	40	22	27	23
44	10	34	35	42	31	34	37	30	12	30	44	37	35	55	17	25
22	25	28	25	33	37	46	18	19	32	42	6	30	27	12	29	23
47	37	28	50	28	27	23	32	24	19	21	23	26	39	53	38	38
39	30	27	20	27	30	45	27	28	29	27	21	41	51	20	31	38
32	18	36	25	27	33	48	25	32	26	38	42	29	19	22	49	25
17	27	33	20	35	34	34	26	31	34	15	28	30	31	14	28	26
35	18	20	26	16	27	23	34	21	30	39	19	28	25	14	21	31
22	21	45	18	32	36	36	28	39	41	32	38	24	38	36	19	31
31	33	27	19	43	31	22	6	33	58	32	30	21	35	26	38	33
29	29	49	30	41	27	38	47	33	23	24	36	21	44	35	53	32
23	47	44	26	51	45	30	30	33	20	31	51	31	31	43	19	35
35	30	26	20	36	35	24	43	31	15	19	36	39	8	44	23	28
39	30	27	46	38	46	27	20	42	30	28	27	29	23	50	15	30
38	23	27	21	22	50	11	39	36	21	30	25	26	12	22	26	38
45	34	36	45	23	16	38	17	36	27	34	32	25	38	35	12	
26	47	43	18	38	45	36	37	30	25	20	24	25	23	26	28	
42	37	27	32	20	15	26	42	26	39	29	37	44	34	50	18	
44	31	30	30	31	36	33	34	20	10	33	42	43	35	45	28	
38	44	26	24	48	37	35	34	34	22	19	34	32	29	32	22	
49	27	36	26	24	1	25	21	28	18	7	27	30	32	35	43	
46	26	26	32	25	45	31	39	20	20	34	27	19	28	34	16	
18	38	40	36	19	25	13	27	36	17	23	36	25	29	24	54	
15	24	10	28	37	20	24	27	27	14	31	34	27	29	35	38	
33	20	21	32	36	22	26	14	28	28	33	36	22	34	29	38	

The Sampling Distribution of the Mean. The first sampling experiment to be described deals with the arithmetic mean. To illustrate the way in which random sampling errors arise, we set up a normal population of values of some character, say X, whose mean is taken as known to be 30 and whose standard deviation is 10; that is, $\mu = 30$, $\sigma = 10$.[1] Sam-

[1] It is conventional to speak of the true values of the population as parameters and to denote them by Greek letters. Correspondingly, Roman letters are the symbols used for the estimates made of parameters or population values from samples.

ples of 5($n = 5$) were chosen at random from the population. By this method, 100 samples of 5 for the variable X were obtained. The individual values for each of the 5 members of each sample are recorded for the 100 samples in Table 2. Note the range in values in the respective samples. For example, in one sample the range in the X-values is from 5 to 47; in another, from 25 to 33.[2]

Next we computed the mean of each of the 100 samples. These values are recorded in Table 3. The 100 means vary between 19.4 and 40.6, and both the highest and lowest means differ from the population mean of 30 by 10.6. Obviously, the means are much less scattered than are the individual values. These fluctuations in mean values are known as *sampling errors.* The amount of sampling error in each mean is the difference between it and the population value, that is, 30. The biggest error with which any one sample of 5 estimates the sample mean is 10.6. The smallest error is found to be 0.2: for instance, the difference between 29.8 and 30.0. None of the 100 estimates is without sampling error.

TABLE 3

THE 100 MEAN VALUES OF THE 100 SAMPLES OF 5 RECORDED IN TABLE 2

23.0	38.6	21.6	34.8	33.6	31.2	29.8	33.8	29.6	23.2
36.8	27.0	28.6	29.0	23.8	35.0	30.6	32.0	23.6	28.4
27.4	28.4	29.8	32.2	28.0	29.2	31.2	34.6	31.8	31.0
32.8	27.6	30.8	40.6	32.0	27.8	26.4	38.0	30.4	27.2
39.0	32.2	27.2	29.8	23.8	21.8	27.0	26.4	28.2	21.6
32.2	34.6	33.2	22.6	35.8	22.6	25.6	29.2	32.4	37.8
31.8	32.4	30.6	28.8	32.8	37.8	28.0	33.8	22.4	23.8
25.6	26.6	37.6	31.4	25.6	21.8	24.0	24.6	38.8	29.4
23.4	31.6	32.0	32.6	32.2	24.6	31.4	24.2	37.6	29.2
31.8	31.8	28.2	26.0	26.0	19.4	35.0	38.4	31.4	32.6

The small samples of 5 give sampling errors greater than would larger samples. Had we taken samples of 50, for instance, the means would have been less scattered, indicating smaller sampling errors. This tendency toward less variation among sampling means and correspondingly smaller differences between sample means and the true mean, and thus a smaller sampling error, would continue as the size of the sample became larger and larger. For example, by calculating the mean of the 100 sample means, we obtain the mean of a single sample of 500 equal to 29.8, a value very close to the population mean of 30.0.

For a sample of a given size, the errors of random sampling increase as the variation among individuals in the population becomes greater.

For example, the estimated mean is $\bar{x}$; the estimated standard deviation, s. This convention is followed throughout this book.

[2] Mahalanobis (see Ref. 5) and others have given tables of random samples from a normal distribution and have shown how to use these tables to get samples of any size for any mean and standard deviation. We have followed this method in several of the empirical sampling experiments described.

In fact, the sampling errors are directly proportional to the increase of variation in the population. As an extreme case, it is obvious that, had there been no variation among individuals in the population sampled and had they all been 30, the means irrespective of sample size would have been 30. Hence, there would have been no sampling errors.

The means in Table 3 may be arranged into a frequency distribution, thus showing the number or frequency of means falling between limits as noted on the base scale (Table 4). This frequency distribution of means is presented in the form of a histogram in Fig. 1. Measures of central location and variability for this frequency distribution of means, which is called the *sampling distribution of means*, can be calculated. It is to be expected that the mean of the 100 means should be the same as the mean of the population being sampled. In our case, the mean of the distribution is found to be 29.8, which agrees closely with 30, the true mean. By increasing the size of the samples, the observed value would become almost exactly 30. The standard deviation of the sampling distribution of means gives an estimate of the size of sampling errors, thus summing up the information concerning the whole distribution of errors. If the standard deviation of the sampling distribution is large, the errors of sampling are, as a whole, large. Correspondingly, if the standard deviation is small, the errors are small. The standard deviation of the frequency distribution of means in Table 4, calculated in the usual manner, is 4.82.

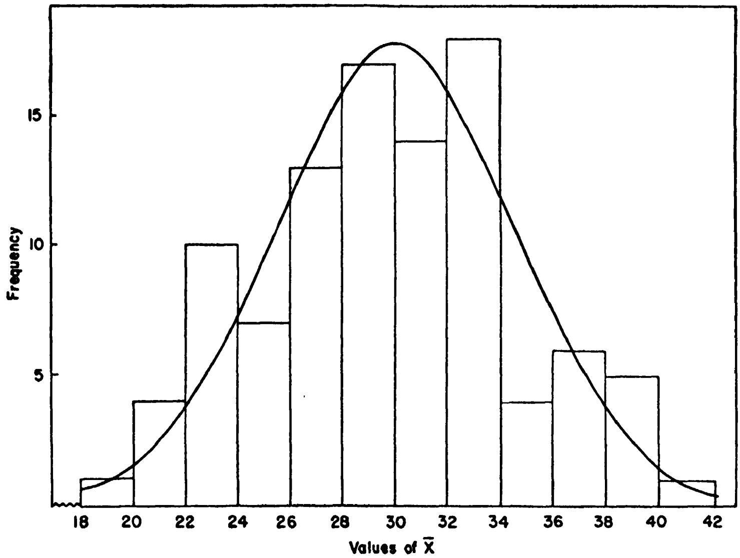


Figure 1. Distribution of means, $\bar{X}$'s of 100 random samples of 5 from a normal population with mean 30 and standard deviation 10. Normal curve superimposed upon the histogram.

As was pointed out earlier, the statistician does not usually carry out this very simple and tedious sampling procedure, since application of the mathematical theory of probability enables him to determine theoretically the sampling distribution and the standard error of a statistic. The application of known mathematical laws gives results that are as accurate as a sampling experiment using millions of samples. Hence, the method of mathematical deduction is at the same time less laborious and more accurate. If the samples are all drawn from a normal population under ideal random sampling conditions, it is known, as in our case, that the sample means are normally distributed about the population mean with a standard deviation equal to $\sigma/\sqrt{n}$, where σ denotes the value of the population standard deviation and n, the number of sampling units. Even if the variable is not normally distributed in the population, it is known that the distribution of totals, or of the means, tends toward normality as the size of the sample is increased.

TABLE 4

FREQUENCY DISTRIBUTION OF THE 100 MEAN VALUES FOR THE SAMPLES OF 5 GIVEN IN TABLE 3 AND THE TEST OF GOODNESS OF FIT

Class interval X	f_0	f_t	f_0	f_t	$(f_0 - f_t)^2$	$\frac{(f_0 - t_t)^2}{f_t}$
41.95 to $+\infty$	0	0.377	6	3.77	4.9729	1.319
39.95 to 41.95	1	0.927				
37.95 to 39.95	5	2.466				
35.95 to 37.95	5	5.405	5	5.41	.1681	0.031
33.95 to 35.95	6	9.687	6	9.69	13.6161	1.405
31.95 to 33.95	17	14.280	17	14.28	7.3984	0.518
29.95 to 31.95	15	17.297	15	17.30	5.2900	0.306
27.95 to 29.95	17	17.213	17	17.21	.0441	0.003
25.95 to 27.95	12	14.101	12	14.10	4.4100	0.313
23.95 to 25.95	7	9.444	7	9.44	5.9536	0.631
21.95 to 23.95	10	5.210	10	5.21	22.9441	4.404
19.95 to 21.95	4	2.361	5	3.59	1.9881	0.554
17.95 to 19.95	1	0.879				
$-\infty$ to 17.95	0	0.353				
Total	100	100.000	100	100.00	$\chi_0^2 = 9.484$	

9 d.f.; $P > .35$

In our case, then, for samples of 5 from the known normal population, we expect the sample means to be normally distributed about 30 with a standard deviation of $\sigma/\sqrt{n} = (10)/\sqrt{5} = 4.472$. Our empirical results, that is, $\bar{x} = 29.8$ and $s_{\bar{x}} = 4.82$, seem to be in close agreement. It is also noted (Table 4) that the observed distribution of means seems to agree very well with the theoretical values educed on the above theory using the mean and standard deviation calculated from the population values. Even with samples as small as 5, the agreement between observation and expectation seems close. We need, of course, a more careful definition of what is meant by "seemingly close." This is given by the chi-square (χ^2) test for goodness of fit.[3] Referring to the χ^2-table (Table III, Appendix) with 9 degrees of freedom we note that for a $\chi_0^2 = 9.484 \sim P > .35$. Therefore, we conclude that we accept the hypothesis that our 100 mean values ($\bar{X}$'s) are normally distributed.

The Sampling Distribution of the Difference Between Means. Our second sampling experiment deals with the differences between the means of random samples. Here we have taken 100 random samples of size 5 for a pair of variables, say X_1 and X_2, which are independent of each other. We know that our parent population of X is normally distributed with mean $\mu = 30$ and standard deviation $\sigma = 10$.

The mean-difference values of X_1 and X_2 for the 100 samples have been calculated and recorded in Table 5.

TABLE 5

THE MEAN-DIFFERENCE VALUES OF X_1 AND X_2 FOR THE 100 SAMPLES OF 5

−4.8	6.4	−12.6	0.6	12.6	−1.4	0	4.6	−0.6	−1.4
15.0	0.2	4.6	−3.4	−0.4	−2.8	0.4	−1.6	0.6	−8.4
−3.4	−1.4	−2.8	−1.0	−1.0	−0.8	1.4	3.6	5.0	5.2
0.6	0.8	−3.0	13.0	4.0	−15.8	−2.8	10.6	1.4	3.6
7.0	−4.2	−3.6	0.2	−11.6	−17.4	−6.4	0.4	0.6	−9.2
−0.4	4.0	6.2	4.4	16.4	−3.2	−12.8	−4.2	−2.2	12.0
1.4	7.4	0.2	−3.4	4.0	8.0	9.4	−6.0	−1.0	−6.0
−4.2	−5.6	13.8	5.4	1.6	−7.4	−6.2	−5.2	7.0	1.2
1.2	3.8	13.0	5.4	3.4	−2.8	5.2	−2.0	14.2	−1.4
1.0	−4.0	−6.2	−7.4	−5.6	−17.6	3.0	2.8	8.0	0.6

From sampling theory it is known that the mean-difference values, $(\bar{X}_2 - \bar{X}_1)$'s, are normally distributed about a mean of 0 with standard deviation $\sqrt{2}\sigma/\sqrt{n}$, where σ is the population standard deviation and n denotes the sample size.

The mean-difference values in Table 5 are arranged in a frequency distribution in Table 6. We find the mean of the mean-difference values to be $\bar{d} = .39$, and the standard deviation, or standard error, of the mean of differences to be $s_{\bar{d}} = 6.736$. The corresponding parameter values are $\mu = 0$ and $\sigma_{\bar{d}} = 6.325$. The observed values are well within the limits of sampling error.

Again we wish to test the goodness of fit of the normal distribution. The theoretical frequencies (f_t) were calculated and are given in Table 6. Chi-square is the appropriate test of goodness of fit of the theoretical and observed distributions. Its value is found to be 10.985. We enter the χ^2-table (Table III, Appendix) with 9 degrees of freedom and $\chi_0^2 = 10.985$. The corresponding probability value is $P > .27$. Hence,

[3] See page 96.

we may conclude that the 100 mean-difference values are normally distributed in accordance with sampling theory.

The Sampling Distribution of the Variance. Our third sampling experiment deals with the variance. Samples of 5 were chosen at random from a normal population with mean $\mu = 30$ and variance $\sigma^2 = 100$. The sum of the squares of the deviations of the observational values X_{ij} from their mean $\bar{X}_i$ for each of the 100 samples was obtained, and these values are recorded in Table 7.

TABLE 6
FREQUENCY DISTRIBUTION OF THE 100 MEAN-DIFFERENCE VALUES FOR THE SAMPLES OF 5 GIVEN IN TABLE 5 AND THE TEST OF GOODNESS OF FIT

Class interval $\bar{X}_2 - \bar{X}_1$	f_0	f_t	$f_0 - f_t$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
17.95 to ∞	0				
15.95 to 17.95	1				
13.95 to 15.95	2 } 9	5.79	3.21	10.3041	1.780
11.95 to 13.95	5				
9.95 to 11.95	1				
7.95 to 9.95	3 } 8	11.55	−3.55	12.6025	1.091
5.95 to 7.95	5				
3.95 to 5.95	11	9.26	1.74	3.0276	0.327
1.95 to 3.95	6	11.31	−5.31	28.1961	2.493
− 0.05 to 1.95	19	12.41	6.59	43.4281	3.499
− 2.05 to − 0.05	13	12.38	0.62	.3844	0.031
− 4.05 to − 2.05	12	11.19	0.81	.8472	0.076
− 6.05 to − 4.05	9	9.18	−0.18	.0324	0.004
− 8.05 to − 6.05	5 } 7	11.33	−4.33	18.7489	1.655
−10.05 to − 8.05	2				
−12.05 to −10.05	1				
−14.05 to −12.05	2				
−16.05 to −14.05	1 } 6	5.60	0.40	.6600	0.029
−18.05 to −16.05	2				
− ∞ to −18.05	0				
Total	100	100.00	0.00	$\chi_0^2 = 10.985$	

d.f. = 9; .30 > P > .20

We have calculated the variance of each sample by dividing the sum of the squares of the deviations of the observation values X_{ij} from $\bar{X}_i$ by $n = 5$. These estimates are called Pearsonian. Thus,

$$s^2_{i_{(P)}} = \frac{\sum_j (X_{ij} - \bar{X}_i)^2}{5} \qquad \begin{pmatrix} i = 1, \cdots, 100 \\ j = 1, \cdots, 5 \end{pmatrix} \qquad (3.01)$$

where $\bar{X}_i$ is the mean of X for the ith sample and X_{ij} is the jth individual in the ith sample.

The $s^2_{i_{(P)}}$ were arranged into a frequency distribution and the theoretical frequencies calculated.

We found the mean and the standard deviation of the $s^2_{i_{(P)}}$ to be as follows:

$$\overline{s^2_{(P)}} = 68.945, \qquad s_{s^2_{(P)}} = 53.279$$

The test of the goodness of fit of the chi-square function gave a $\chi^2_0 = 11.836$. Entering the table of χ^2 (Table III, Appendix) with 6 d.f., the probability was found to be $.10 > P > .05$. It was concluded that the sampling distribution of the $s^2_{i_{(P)}}$ follows the χ^2 distribution.

TABLE 7

THE 100 VALUES OF THE SUMS OF SQUARES BASED ON SAMPLES OF 5 FROM A NORMAL POPULATION WITH MEAN 30 AND STANDARD DEVIATION 10

112.0	53.2	85.2	318.8	235.2
402.8	443.2	174.8	241.2	849.2
229.2	408.8	186.0	370.8	138.8
180.8	60.8	442.0	97.2	25.2
240.0	66.8	444.8	202.0	284.8
970.8	150.8	470.8	507.2	1613.2
362.8	401.2	354.8	214.0	339.2
393.2	437.2	330.2	738.0	466.8
169.2	518.0	1158.8	323.2	381.2
322.8	250.8	82.0	422.0	93.2
181.2	584.8	168.8	176.8	218.8
180.0	154.0	74.0	86.0	231.2
93.2	46.8	164.8	313.2	626.0
85.2	353.2	128.8	542.0	900.8
354.8	730.8	234.8	57.2	187.2
485.2	981.2	257.2	176.8	764.8
201.2	470.8	632.8	346.8	400.8
575.2	977.2	118.8	73.2	249.2
439.2	455.2	433.2	228.8	48.8
534.8	390.0	111.2	303.2	63.2

Each of the $s^2_{i_{(P)}}$ is an estimate of the variance, σ^2, of the population (= 100) from which we were sampling. Therefore, we expect the mean of the 100 samples of 5 to be approximately equal to σ^2 or 100. From sampling theory it is known that the expected standard deviation of the $s^2_{(P)}$'s is:

$$\sigma_{s^2_{(P)}} = \frac{\sigma^2\sqrt{2(n-1)}}{n} = 100\frac{\sqrt{8}}{5} = 56.57 \tag{3.02}$$

Our calculated value of the standard deviation of the 100 obtained values of $s^2_{i_{(P)}}$ was 53.279. Thus, both the mean and the standard deviation of the 100 sample values of $s^2_{i_{(P)}}$ differ considerably from expectation. Estimates are considered biased estimates if in repeated sampling their mean or mathematical expectation does not equal the true, or population, value. Our obtained mean value, $\overline{s^2_{(P)}}$ is too low. It is 1.22 standard errors below the true value. Although the obtained mean is within the limit of random sampling fluctuations, we shall consider whether or not a closer agreement with expectation can be obtained.

We shall now calculate our estimate in a different manner. Define:

$$s^2_{i_{(u)}} = \frac{\sum_j (X_{ij} - \bar{X}_i)^2}{4} \quad \begin{pmatrix} i = 1, \cdots, 100 \\ j = 1, \cdots, 5 \end{pmatrix} \tag{3.03}$$

where the subscript (u) indicates the unbiased estimate.

We calculated the 100 values of $s^2_{i_{(u)}}$. They are recorded in Table 8.

TABLE 8

100 Unbiased Estimates, $s^2_{i_{(u)}}$ Calculated from the Sums of Squares in Table 7

28.0	45.3	13.3	146.2	21.3	42.2	79.7	44.2	58.8	54.7
100.7	45.0	110.8	38.5	43.7	18.5	60.3	21.5	212.3	57.8
57.3	23.3	102.2	11.7	46.5	41.2	92.7	78.3	34.7	156.5
45.2	21.3	15.2	88.3	110.5	32.2	24.3	135.5	6.3	225.2
60.0	88.7	16.7	182.7	111.2	58.7	50.5	14.3	71.2	46.8
242.7	121.3	37.7	245.3	117.7	64.3	126.8	44.2	403.3	191.2
90.7	50.3	100.3	117.7	88.7	158.2	53.5	86.7	84.8	100.2
98.3	143.8	109.3	244.3	84.8	29.7	184.5	18.3	116.7	62.3
42.3	109.8	129.5	113.8	289.7	108.3	80.8	57.2	95.3	12.2
80.7	133.7	62.7	97.5	20.5	27.8	105.5	75.8	23.3	15.8

Define again:

$$\overline{s^2_{(u)}} = \frac{\sum_i s^2_{i_{(u)}}}{100} \qquad (i = 1, \cdots, 100) \tag{3.04}$$

$$s_{s^2_{(u)}} = \sqrt{\frac{\sum_i (s^2_{i_{(u)}} - \overline{s^2_{(u)}})^2}{100}} \qquad (i = 1, \cdots, 100) \tag{3.05}$$

where the subscript (u) again indicates the unbiased estimate.

From Table 8 we obtain

$$\overline{s^2_{(u)}} = 85.92, \qquad s_{s^2_{(u)}} = 67.04$$

Theoretically, we have

$$\mu_{s^2_{(u)}} = \sigma^2 = 100 \tag{3.06}$$

$$\sigma_{s^2_{(u)}} = \sigma^2 \sqrt{\frac{2}{n-1}} = 100\sqrt{\frac{2}{4}} = 70.71 \tag{3.07}$$

We now observe that our calculated value, $\overline{s^2_{(u)}} = 85.92$, is .46 standard error below the true value, well within the limits of random sampling fluctuations.

We may now state that the usual method of calculating the estimate of the population variance, that is, $s^2_{i_{(P)}}$ as an estimate of σ^2, gives a biased estimate. The explanation of the bias is that, if we use $s^2_{(P)}$, its mean in repeated sampling is not σ^2 but $\frac{n-1}{n}\sigma^2$, where n is the size of the sample. In our case:

$$\overline{s^2_{(P)}} \text{ is an estimate of } \sigma^2 \frac{(n-1)}{n} \qquad \text{or} \qquad 100 \cdot \frac{4}{5} = 80 \qquad (3.08)$$

Now, our calculated value of $\overline{s^2_{(P)}} = 68.95$ does not differ significantly from the theoretical value 80; that is, it is .44 standard deviation below, and the difference is due to sampling error alone.

The amount of the bias in using $s^2_{(P)}$ as an estimate of σ^2 is evidently $\frac{1}{n}\sigma^2$. The theoretical standard deviation of the distribution of $s^2_{(P)}$, when n is large, is $\sigma^2\sqrt{2(n-1)}/n$. It may be worth noting the relative magnitude of the bias and sampling error. The respective values for various sizes of samples are recorded in Table 9.

The values in columns (2) and (3) of Table 9 show that the bias is substantial in comparison with random sampling error, especially for small samples, for instance $n = 50$ or less. The conclusion is that, since there is no justification for willfully introducing a bias, the unbiased estimate, $s^2_{(u)}$, should be used when estimates of the population variance are required as in problems of statistical inference. When mere description is involved, $s^2_{(P)}$, may properly be used.

TABLE 9

COMPARISON OF THE BIAS IN USING $s_{(P)}{}^2$ AS AN ESTIMATE OF σ^2 WITH SAMPLING ERRORS

Size of sample (n)	Relative amount of bias $\left(\frac{1}{n}\right)$	Relative amount of sampling error $\frac{\sqrt{2(n-1)}}{n}$
(1)	(2)	(3)
2	0.50	0.71
3	0.33	0.67
5	0.20	0.57
10	0.10	0.42
20	0.05	0.31
50	0.02	0.20
100	0.01	0.14

Likewise, it is customary to consider the square root of the unbiased estimate of the variance as the unbiased estimate of the standard deviation; that is,

$$s_{i_{(u)}} = \sqrt{s^2_{i_{(u)}}} = \sqrt{\frac{\sum_j (X_{ij} - \bar{X}_i)^2}{n-1}} \quad \begin{pmatrix} i = 1, \cdots, 100 \\ j = 1, \cdots, 5 \end{pmatrix} \tag{3.09}$$

It may be stated, however, that since $\overline{s^2_{(u)}}$ in repeated sampling equals σ^2, Equation (3.09) does not imply that in repeated sampling the mean, $(\bar{s}_{(u)}) = \sigma$. It is well known that the mean of a sum of numbers does not exactly equal the square root of the arithmetic mean of their squares. To illustrate:

$$\bar{X} = \tfrac{1}{6}(1 + 3 + 5 + 6 + 8 + 9) = 5\tfrac{1}{3}$$

$$\sqrt{\text{Mean of } (X^2)} = \sqrt{\tfrac{1}{6}(1 + 9 + 25 + 36 + 64 + 81)} = 6$$

We have arranged the hundred $s^2_{(u)}$'s in Table 8 into a frequency distribution (Table 10). The theoretical values have been also calculated and the theoretical curve based on these has been constructed in Fig. 2. The test of the goodness of fit of the theoretical for the observed frequencies gives a χ_0^2 value of 15.25 with $P > .05$. The model, in this case built up of the sampling errors of the variance, is known as the *chi-square curve.* This model is important in statistical theory and practice (see Table III, Appendix).

TABLE 10
FREQUENCY DISTRIBUTION OF THE 100 UNBIASED ESTIMATES $s^2_{(u)}$ OF THE POPULATION VARIANCE IN TABLE 8 AND THE TEST OF GOODNESS OF FIT

Interval	f_0	f_t	$f_0 - f_t$	$\frac{(f_0 - f_t)^2}{f_t}$
331.925 to ∞ 291.700 to 331.925 237.200 to 291.700 194.475 to 237.200	7	10	−3	0.90
149.725 to 194.475	5	10	−5	2.50
121.950 to 149.725	6	10	−4	1.60
83.925 to 121.950	27	20	7	2.45
54.875 to 83.925	16	20	−4	0.80
41.225 to 54.875	14	10	4	1.60
26.600 to 41.225	8	10	−2	0.40
17.775 to 26.600	9	5	4	3.20
10.725 to 17.775 7.425 to 10.725 0.000 to 7.425	8	5	3	1.80
Total	100	100	0	χ_0^2 = 15.25

d.f. = 8; .10 > P > .05

The Sampling Distribution of *t*. We have now considered two principal models, the normal and the chi-square, which the statistician has developed. It is to be remembered that in the development of both of these models it was assumed that the variance or standard deviation of the population was known. It is not often the case, however, in experimental work that the population value is known. Furthermore, the experimenter is usually dealing with small samples. The construction of a model against which experimental results of this kind could be compared would indeed be a genuine contribution to the research worker. Let us trace the way in which this problem was solved.

Since the population standard deviation is unknown, the only source of information concerning it is that provided by the sample. It was observed (see Table 8) that the sample variance, and hence the standard deviation, is often very different from the population standard deviation. Each of these variances or standard deviations was an estimate of the population value. The smallest standard deviation was $\sqrt{6.3}$ or 2.51; the largest, $\sqrt{403.30}$ or 20.08. It was essential that a model to be effective should take these sampling fluctuations into account. How this was done was to set up a ratio of the difference between the sample mean and population mean to its estimated standard error. This ratio was called t. In mathematical terms we may proceed as follows:

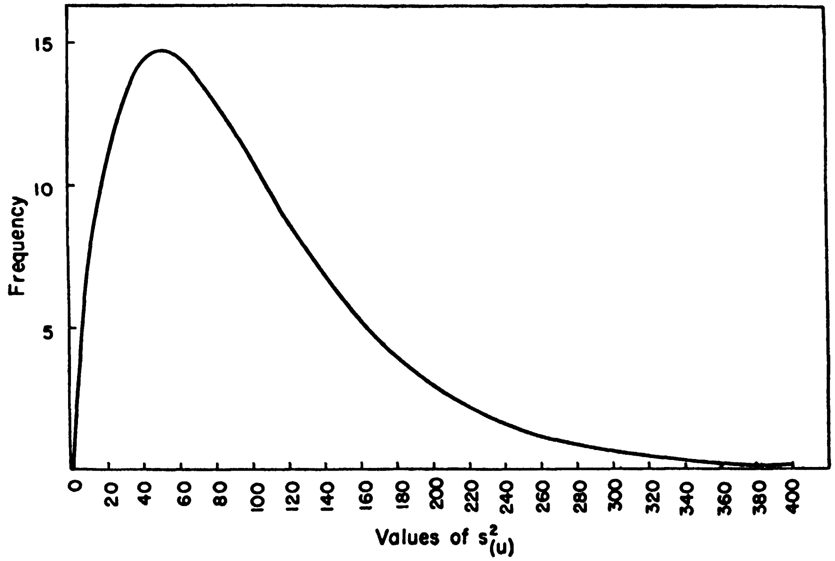


Figure 2. The Chi-square distribution curve based on the unbiased estimates of the variance, $s^2_{(u)}$'s, of 100 random samples of 5 from a normal population with mean 30 and variance 100 (Table 10).

Suppose that the parameter σ is unknown, though $\mu = 30$ in our parent population (μ may or may not be known). Define:

$$t_i = \frac{(\bar{X}_i - \mu)}{\sqrt{\dfrac{\sum\limits_j (X_{ij} - \bar{X}_i)^2}{n(n-1)}}} = \frac{(\bar{X}_i - \mu)\sqrt{n(n-1)}}{\sqrt{\sum\limits_j (X_{ij} - \bar{X}_i)^2}} \begin{pmatrix} i = 1, \cdots, 100 \\ j = 1, \cdots, \ \ 5 \end{pmatrix} \qquad (3.10)$$

where $\bar{X}_i$ is the mean value of the ith sample and X_{ij} is the jth individual in the ith sample. Then

$$y = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m}\sqrt{\pi}}\left[1 + \frac{t_i^2}{m}\right]^{-\frac{m+1}{2}} \tag{3.11}$$

where y is the ordinate value for a specific value of t and m is the number of degrees of freedom; Γ denotes a Gamma function.

In our sampling experiment we took 100 samples of size 5,

$$\mu = 30, \qquad n = 5, \qquad m = 4$$

The 100 t_i-values were calculated and these were recorded in Table 11.

TABLE 11

THE 100 t-VALUES FOR 100 RANDOM SAMPLES OF 5 FROM A POPULATION WITH MEAN OF 30 AND UNKNOWN VARIANCE σ^2

−2.9580	−5.1504	1.7442	−0.0501	−0.1166
1.5152	−0.2974	−2.0972	0.1728	−0.9822
−0.7680	−0.0442	−0.6558	0.2787	0.6833
0.9313	0.4588	0.4254	−1.6330	0.3563
2.5981	−1.5321	−1.3147	−0.9440	−0.4770
0.3158	1.1654	1.1954	−0.8737	0.2672
0.4226	0.1340	0.6648	−0.6114	−1.8454
−0.9923	1.6255	−1.0684	−0.9877	1.8215
−2.2691	0.3930	0.2890	0.3483	1.7408
0.4480	−0.5083	−1.9755	1.0885	0.6485
2.8572	0.8877	0.4131	1.2781	−2.0559
−1.0000	−0.3604	2.5994	0.9645	0.4706
−0.7412	1.4382	−0.2787	1.1624	0.1787
−1.1628	2.5224	−0.8669	1.5368	−0.4172
0.5223	−0.0331	−2.3932	−2.1287	−2.7456
0.9339	−1.0565	−2.0635	−0.2691	1.2614
0.7567	−0.2473	1.3867	0.9126	−1.3850
−0.6340	0.2003	−3.3645	−2.8226	−0.1700
0.3414	0.5450	−1.1603	−1.7148	−0.5121
0.3481	−0.9058	−4.4954	2.1574	1.4626

Theoretically, we have

$$\mu_t = 0$$

$$\sigma_t = \sqrt{\frac{m}{m-2}} = 1.414$$

where μ_t and σ_t are the mean and standard deviation of all the possible

t-values, respectively. For our 100 t-values we have

$$\bar{t} = \frac{\sum\limits_i t_i}{100} = -.165$$

$$s_t = \sqrt{\frac{\sum\limits_i (t_i - \bar{t})^2}{100}} = 1.490 \qquad (i = 1, \cdots, 100)$$

Again, we wish to test the goodness of fit for the t-distribution by using the χ^2-criterion. This test is given in Table 12.

TABLE 12

DISTRIBUTION OF THE 100t-VALUES FROM MEANS OF SAMPLES OF 5 AND TEST OF GOODNESS OF FIT

Class interval of t	f_0	f_t	$f_0 - f_t$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
4.005 to $+\infty$	0	0.79			
3.005 to 4.005	0 } 5	1.20 } 5.78	*−0.78*	0.6084	0.105
2.005 to 3.005	5	3.79			
1.005 to 2.005	15	12.78	2.22	4.9284	0.386
0.005 to 1.005	30	31.25	*−1.25*	1.5625	0.050
−0.995 to 0.005	26	31.36	*−5.36*	28.7296	0.916
−1.995 to *−0.995*	12	13.00	*−1.00*	1.0000	0.077
−2.995 to *−1.995*	9	3.83			
−3.995 to *−2.995*	1 } 12	1.19 } 5.83	6.17	38.0689	6.530
$-\infty$ to *−3.995*	2	0.81			
Total	100	100.00	0.00	$\chi_0^2 = 8.064$	

d.f. = 5; $P > .14$

Referring to the χ^2 table (Table III, Appendix) with $\chi_0^2 = 8.064$ and 5 degrees of freedom, we find that $P > .14$. Therefore, we conclude that our 100 t-values are distributed as the t-function.

We have arranged the 100 t-values in a frequency distribution and plotted the histogram. The theoretical frequency distribution of t has been calculated and the corresponding curve has been superimposed on the histogram (Fig. 3). The theoretical frequency curve of the sampling distribution of t is a symmetrical leptokurtic curve. Tables (see Table II, Appendix) have been prepared which enable one to determine for a given size of sample the probability of getting a value of t greater than or equal to $\pm t_0$, or the value in the sample, due to random sampling errors alone in repeated sampling. Against this model, when it is appropriate for the problem involved, the experimenter may then compare his experimental results with the view of examining their significance.

Contribution of "Student." It is fitting here to point out the significance of the contribution of the writer who signed himself "Student" to the refinement of the classical theory of errors. First, it is usually held that the date of his publication (Ref. 7), 1908, is the beginning of modern statistical theory and practice. When Student began his work as one of the brewers of Guinness, Son and Company, the available statistical tools were postulated upon large sampling theory. In the course of his work it was necessary for him to draw conclusions from the results of small samples which themselves furnished the only indication of their variability. Rigorous conclusions under such conditions became possible through Student's determination of the exact sampling distribution of the statistic, thus making allowance for its sampling errors. He demonstrated that notwithstanding these sampling errors, which in the case of very small samples are large, it was possible to derive a test of significance both rigorous and exact. Since the number of degrees of freedom is one of the parameters in the equation of the sampling distribution, the restriction previously set up, namely, that the sample must be "large," was removed.

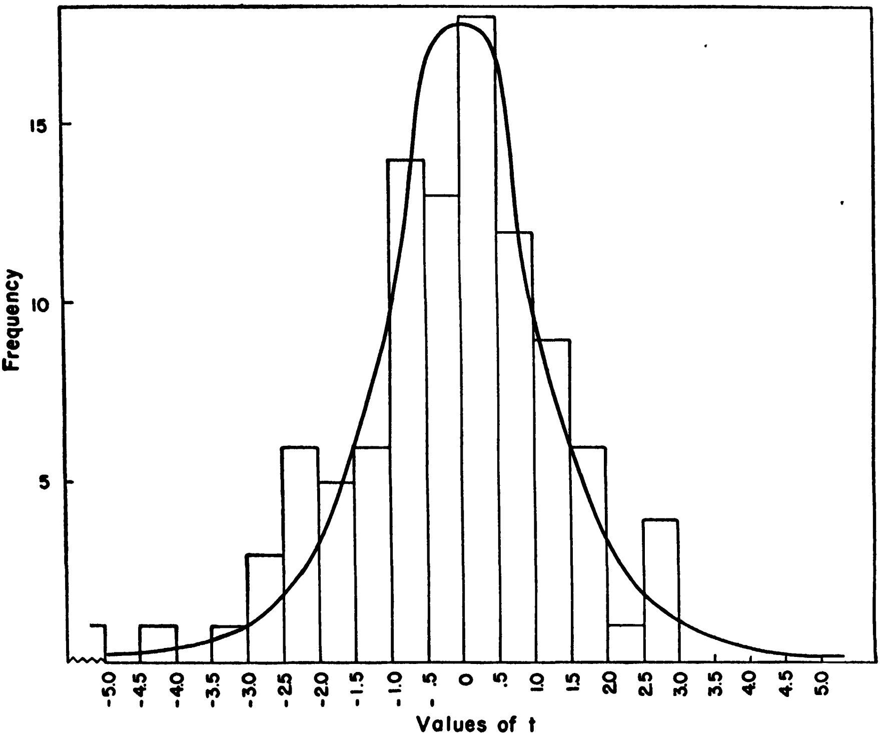


Figure 3. Distribution of the t-values of 100 random samples of 5. Theoretical curve of the t-distribution superimposed upon the histogram.

The applicability of Student's test has, of course, been greatly extended by modern research in mathematical statistics. There are two memorials to Student which will undoubtedly endure: (1) a "Studentized" function, that is, a statistic whose sampling distribution, originally involving the standard deviation of the population, is altered so that its sampling distribution uses quantities calculated only from the sample; (2) an exact test of significance, that is, a test which depends on a known probability distribution and thus is independent of irrelevant unknown parameters.

The t-Distribution of the Difference between Means. More frequent than the need of comparing experimental results of a single sample with a model is that of comparing the results for two independent samples, for instance, the difference between the means of the experimental and control groups. Therefore, we present the results of an empirical sampling experiment dealing with the model built up of the sampling errors of differences between means. The samples used in this case were those obtained in the sampling experiment described on page 37 and in Table 5, where 100 random samples of five for a pair of variables X_1 and X_2, which are independent of each other, were taken. Here we have assumed that we do not know the population standard deviation and hence have to estimate it from the sample. The results in this case are found to be described by the model t-distribution. Suppose that we do not know the parameters, σ_1 and σ_2, though we know $\mu_1 = 30$ and $\mu_2 = 30$ in our parent population (μ_1 and μ_2 may or may not be known). Define:

$$t_i = \frac{(\bar{X}_{1i} - \bar{X}_{2i})}{\sqrt{\sum_j (X_{1ij} - \bar{X}_{1i})^2 + \sum_j (X_{2ij} - \bar{X}_{2i})^2}} \cdot \frac{\sqrt{n_1 + n_2 - 2}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \begin{pmatrix} i = 1, \cdots, 100 \\ j = 1, \cdots, 5 \end{pmatrix} \qquad (3.12)$$

where $\bar{X}_{1i}$ is the mean value of X_1 in the ith sample; $\bar{X}_{2i}$ is the mean value of X_2 in the ith sample; X_{1ij} is the jth individual in the ith sample for X_1; and X_{2ij} is the jth individual in the ith sample for X_2. Then it may be shown (Ref. 1) that for samples, n_1 and n_2 from a normal population, the distribution of t is given by

$$y = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\sqrt{m}\sqrt{\pi}}\left[1 + \frac{t_i^2}{m}\right]^{-\frac{m+1}{2}} \qquad (3.13)$$

where y is the ordinate value for a specific value of t and m is the number of degrees of freedom. In our case,

$$n_1 = 5, \qquad n_2 = 5, \qquad m = n_1 + n_2 - 2 = 8$$

The t-values have been calculated for the 100 samples of 5 for a pair of values X_1 and X_2 and are recorded in Table 13.

Theoretically, we have

$$\mu_t = 0$$

$$\sigma_t = \sqrt{\frac{m}{m-2}} = 1.1547$$

where μ_t and σ_t are the mean and standard deviation of all the possible t-values, respectively. For our 100 t-values, we have

$$\bar{t} = \frac{\sum_i t_i}{100} = .0234 \qquad (i = 1, \cdots, 100)$$

$$s_t = \sqrt{\frac{\sum_i (t_i - \bar{t})^2}{100}} = 1.2459$$

Finally, we wish to test the goodness of fit for the t-distribution by using the χ^2-criterion. The test is given in Table 14.

TABLE 13

THE 100 t-VALUES FOR MEAN DIFFERENCES OF 100 RANDOM SAMPLES OF 5 FOR PAIRS OF VALUES X_1 AND X_2

−0.9851	*−2.4385*	1.5956	0.0000	*−0.1539*
3.2678	0.6438	*−0.0719*	0.0591	0.0708
−0.5723	*−0.6053*	*−0.2113*	0.2773	1.4153
0.1244	*−0.7650*	0.6870	*−0.3565*	0.3285
1.3229	*−0.5523*	*−1.6920*	*−1.7380*	0.1232
−0.0400	1.7302	3.0993	*−2.1881*	*−0.2028*
0.2387	0.0299	0.7405	1.8521	*−0.1302*
−0.5922	2.1985	0.2455	*−0.7852*	1.0640
0.2519	2.1117	0.3732	0.8348	2.3370
0.0969	*−1.1153*	*−1.2291*	0.3926	3.2338
0.8221	0.0856	*−0.3300*	0.9167	*−0.2725*
0.0358	*−0.8355*	*−0.5859*	*−0.6295*	*−1.7950*
−0.5491	*−0.2123*	*−0.1353*	0.5357	0.6711
0.1451	2.1691	*−4.3324*	1.7070	0.4798
−0.5952	0.0301	*−2.3705*	0.0983	*−1.4368*
0.4930	0.5699	*−0.4897*	*−1.2175*	1.1948
1.8076	*−0.6283*	1.1622	*−0.8290*	*−0.9685*
−0.8922	0.4934	*−1.8228*	*−1.2257*	0.2524
0.5445	0.9909	*−0.5337*	*−0.1298*	*−0.1801*
−0.5213	*−1.5863*	*−2.6307*	0.4127	0.1375

Referring to the χ^2 table (Table III, Appendix) with $\chi_0^2 = 8.817$ and with 7 degrees of freedom, we find that $P > .25$. Therefore, we conclude that our 100 t-values are distributed as the t-function.

The Sampling Distribution of the Correlation Coefficient. We now present the results of a sampling experiment which illustrate the theoretical or statistical model built up from the sampling errors of the correlation coefficient in repeated random sampling from a population in which the true correlation is known to be zero.

The samples used were the ones obtained by taking 100 samples of 5 pairs of values from a normal population in which there was no correlation at all between the variables (see page 37).

TABLE 14

TEST OF GOODNESS OF FIT OF THE THEORETICAL t-DISTRIBUTION FOR THE OBSERVED t-VALUES OF TABLE 13

Classes	f_0	f_t	$f_0 - f_t$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
3.505 to ∞	0				
3.005 to 3.505	3				
2.505 to 3.005	0 }12	8.57	3.43	11.7649	1.373
2.005 to 2.505	4				
1.505 to 2.005	5				
1.005 to 1.505	5	8.66	*−3.66*	13.3956	1.547
0.505 to 1.005	11	14.11	*−3.11*	9.6721	0.685
0.005 to 0.505	24	18.46	5.54	30.6916	1.663
−0.495 to 0.005	15	18.58	*−3.58*	12.8164	0.690
−0.995 to *−0.495*	18	14.16	3.84	14.7456	1.041
−1.495 to *−0.995*	5	8.77	*−3.77*	14.2129	1.621
−1.995 to *−1.495*	5				
−2.495 to *−1.995*	3				
−2.995 to *−2.495*	1 }10	8.69	1.31	1.7161	0.197
−3.495 to *−2.995*	1				
− ∞ to *−3.495*					
Total	100	100.00	0.00	$\chi_0^2 = 8.817$	

d.f. = 7; $P > .25$

Let us define:

$$r_1 = \frac{\sum_j \{(X_{1ij} - \bar{X}_{1i})(X_{2ij} - \bar{X}_{2i})\}}{\sqrt{\sum_j (X_{1ij} - \bar{X}_{1i})^2 \sum_j (X_{2ij} - \bar{X}_{2i})^2}} \qquad \begin{pmatrix} i = 1, \cdots, 100 \\ j = 1, \cdots, 5 \end{pmatrix} \tag{3.14}$$

where $\bar{X}_{1i}$ and $\bar{X}_{2i}$ are the means for X_1 and X_2, respectively, in the ith sample, and X_{1ij} and X_{2ij} are the jth individual in the ith sample for X_1 and X_2, respectively. The r-values for the 100 samples in Table 5 have been calculated and recorded in Table 15.

Then it may be shown that r is distributed in repeated sampling in the following function:

$$y = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)} (1 - r_i^2)^{\frac{n-4}{2}} \tag{3.15}$$

where y is the ordinate value for a specific value of r and n is the sample size. In our case, $n = 5$.

TABLE 15

THE 100 VALUES OF THE CORRELATION COEFFICIENT r CALCULATED FOR 100 RANDOM SAMPLES OF 5 FROM A POPULATION IN WHICH THE TRUE CORRELATION IS ZERO

−.829	−.954	.359	−.310	.294
−.669	−.134	.349	.876	−.614
−.592	−.862	.168	−.748	.482
.142	−.091	−.235	.663	−.726
.126	−.926	.289	.152	−.640
.548	−.403	−.706	−.274	−.114
.522	.391	−.636	.370	.114
.830	.111	.111	.150	.482
.563	.437	.196	.516	.720
.678	.733	−.212	.007	−.744
−.392	.196	−.108	−.124	−.690
.378	.847	.196	.640	−.885
−.625	.079	−.531	−.589	.665
−.613	.549	−.304	.221	−.763
−.032	−.483	−.294	.216	.737
.528	.368	−.716	−.638	−.238
.481	−.881	−.388	.075	−.832
−.296	−.276	−.485	−.010	−.349
.298	.089	−.232	.725	−.574
−.278	.120	.725	.159	−.473

Theoretically, we have

$$\mu_r = 0$$

$$\sigma_r = \frac{1}{\sqrt{n-1}} = .500$$

where μ_r and σ_r are the mean and the standard deviation of all the possible r-values, respectively. For our 100 r-values, we have

$$\bar{r} = \frac{\sum_i r_i}{100} = -.0596$$

$$s_r = \sqrt{\frac{\sum_i (r_i - \bar{r})^2}{100}} = .5076$$

Finally, we wish to test the goodness of fit for the r-distribution by using the χ^2-criterion.[4] The test is given in Table 16.

[4] F. N. David, *Tables of the Correlation Coefficient, Biometrika* Office, University College, London, 1938.

Referring to the χ^2-table (Table III, Appendix) with $\chi^2 = 12.004$ and with 8 degrees of freedom, we have $P > .14$. Therefore, we conclude that our 100 r-values are distributed as the r-function [Formula (3.15)].

It is known from sampling theory that for large samples, where n is larger than 100, r is approximately normally distributed about zero ($\rho = 0$) with a standard deviation equal to $\dfrac{1}{\sqrt{n-1}}$. In our sampling experiment it is noted that the mean of the 100 sample values of r was $-.0596$ and that the standard deviation was .5076. Even with samples of $n = 5$, these values agree closely with the expected values of 0 and .500. At least for large samples, the normal curve might be used for the mathematical model when the true value of $\rho = 0$, against which the experimental results might be compared. An exact test, however, is available based on the t-distribution as outlined above. In this case,

TABLE 16

DISTRIBUTION OF 100 CORRELATION COEFFICIENTS AND THE TEST OF GOODNESS OF FIT OF THE THEORETICAL FOR OBSERVED VALUES OF r

Class interval	f_0	f_t	$f_0 - f_t$	$(f_0 - f_t)^2$	$\dfrac{(f_0 - f_t)^2}{f_t}$
.805 to 1.000 .605 to .805	$\left.\begin{matrix}3\\8\end{matrix}\right\}11$	$\left.\begin{matrix}5.01\\8.97\end{matrix}\right\}13.98$	−2.98	8.8804	0.635
.405 to .605	10	10.96	−0.96	0.9216	0.084
.205 to .405	11	12.10	−1.10	1.2100	0.100
.005 to .205	18	12.64	5.36	28.7296	2.273
− .195 to .005	6	12.65	−6.65	44.2225	3.496
− .395 to − .195	14	12.14	1.86	3.4596	0.285
− .595 to − .395	8	11.03	−3.03	9.1809	0.832
− .795 to − .595	15	9.10	5.90	34.8100	3.825
−1.000 to − .795	7	5.40	1.60	2.5600	0.474
Total	100	100.00	0.00	$\chi_0^2 = 12.004$	

d.f. = 8; $P > .14$

$$t = \frac{r\sqrt{f}}{\sqrt{1 - r^2}}; \qquad f = n - 2 \tag{3.16}$$

This test is particularly useful for small samples.

When the correlation in the population is not zero, that is, when $\rho \neq 0$, the sampling distribution of r is distributed about ρ with a standard deviation, or standard error approximately equal to $1 - \rho^2/\sqrt{n-1}$. When $\rho = 0$, this reduces to the standard deviation given above, or $1/\sqrt{n-1}$. With large samples and moderate or small values of ρ the sample value r may be substituted for the unavailable ρ, for example, $1 - r^2/\sqrt{n-1}$ as a measure of the sampling error of r. With small samples, however, the sample value, r, often differs greatly from the true value. Furthermore, the sampling distribution departs widely from normality so that the test of significance based upon the formula for large samples may be highly misleading. The constants of distribution of r for samples of $n = 20$ from a normal population as given in Table 17 are illustrative of this point.[5]

TABLE 17
CONSTANTS OF DISTRIBUTION OF r IN SAMPLES ($N = 20$) FROM A NORMAL POPULATION

ρ	0	.2	.4	.6	.8	.9
β_1	0.000	0.066	0.260	.650	1.400	2.060
β_2	2.710	2.820	3.170	3.910	5.420	6.870
σ_r	0.229	0.221	0.197	0.154	0.091	0.049
$\dfrac{1-\rho^2}{\sqrt{N-1}}$	0.229	0.220	0.193	0.147	0.083	0.044

Fisher solved these and related problems by using the transformation

$$\begin{aligned} z' &= \tanh^{-1} r \\ &= \tfrac{1}{2} \log_e \left(\frac{1+r}{1-r}\right) \end{aligned} \tag{3.17}$$

z' is to a first approximation normally distributed about the population value $\xi + \dfrac{\rho}{2(n-1)}$ for all values of ρ with a standard deviation $\sqrt{\dfrac{1}{n-3}}$. The form of the distribution of z' is nearly independent of the value of ρ in the population. The close approximation to normality of the z'-distribution is noted from the constants of distribution of z' given in Table 18.

TABLE 18
CONSTANTS OF DISTRIBUTION OF z' IN SAMPLES ($N = 20$) FROM A NORMAL POPULATION

ρ	Mean $(z' - \xi)$	$\sigma_{z'}$	β_1	β_2
0	.0000	.2423	.0000	3.116
.2	.0053	.2422	.0000	3.117
.6	.0159	.2412	.0000	3.118
.9	.0249	.2398	.0000	3.114

Other Uses of Statistical Models. We have now illustrated how the statistician builds up statistical or mathematical models against which experimental results may be checked with a view to examining their significance. In order for the reader to gain an insight into the process, a

[5] See page 149 for criteria of normality.

series of empirical sampling experiments was presented. The three principal models illustrated were the normal, the t, and the chi-square. Certain other uses of one or another of these models and of models not previously illustrated for the research worker will now be considered.

The Difference between Correlation Coefficients. The research worker is frequently interested in comparing the relative intensity of relationships for different characters. Although an exact test of significance is not available for such purposes, a test based on Fisher's z'-transformation of the correlation coefficient is valuable and sufficiently accurate for most practical problems (Ref. 3). Let

$$z_1' = \tfrac{1}{2} \log_e \frac{1 + r_1}{1 - r_1}$$
$$z_2' = \tfrac{1}{2} \log_e \frac{1 + r_2}{1 - r_2}$$

where r_1 and r_2 are two correlation coefficients calculated from random samples of n_1 and n_2 individuals, respectively.

$z_1' - z_2'$ varies normally about $\frac{\rho}{2}\left(\frac{1}{n_1 - 1} - \frac{1}{n_2 - 1}\right)$ $(\doteqdot 0)$ with standard deviation $\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$.

Therefore, the quantity

$$X = \frac{z_1' - z_2'}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \tag{3.18}$$

may be assumed to be normally distributed about zero with a standard deviation of unity when the true correlation coefficients in the sampled parent population are in fact equal. The sampling distribution, then, of X in repeated sampling may be assumed to be normal, and the experimental result, X_0, may be compared against the normal model.

The Combination of Correlation Coefficients. The z'-transformation is valuable for use in problems involving the averaging of several sample values of r from the same population in order to get the combined estimate of ρ. Thus the weighted arithmetical mean is

$$\bar{z}' = \frac{(n_1 - 3)z_1' + (n_2 - 3)z_2' + \cdots}{(n_1 - 3) + (n_2 - 3) + \cdots} \tag{3.19}$$

and the standard error of $\bar{z}'$ is

$$s_{\bar{z}'} = \frac{1}{\sqrt{(n_1 - 3) + (n_2 - 3) + \cdots}} \tag{3.20}$$

The ratio $X_0 = \bar{z}'/s_{\bar{z}'}$ may then be referred to the normal model to determine the probability that a value as great as or greater than X_0 could be obtained in repeated sampling by random sampling errors alone.

Correlations on the Same Sample. Comparisons are sometimes made among correlation coefficients based on the same sample. Hotelling (Ref. 4) has given the exact solution of the problem of testing the significance of the difference between r_{y1} and r_{y2} under the conditions that the significance is to be interpreted with respect to subpopulations of possible samples for which predictors X_1 and X_2 take the same set of values as those found in the obtained sample. Thus, F, or the variance ratio, is

$$F = \frac{(r_{y1} - r_{y2})^2(N - 3)(1 + r_{12})}{2(1 - r_{12}^2 - r_{y1}^2 - r_{y2}^2 + 2r_{12}r_{y1}r_{y2})} \qquad n_1 = 1; \qquad n_2 = N - 3 \tag{3.21}$$

where r_{y1} is the correlation coefficient between the predictor X_1 and the predictand y; r_{y2}, between X_2 and y; and r_{12} is the correlation between X_1 and X_2.

The assumption underlying the test is that (1) y has the univariate normal distribution for each set of values of X_1 and X_2, independently for the different sets with (2) a common variance σ^2 and (3) linear regression of y on X_1 and X_2, respectively.

Hotelling also developed formulas for determining the selection of (a) one variate from among three or more and (b) additional variates when some have been chosen. His principal solutions of tests of significant differences among $r_{y1}, \ldots, r_{yp}$ are given in Ref. 4.

Fisher's z-Distribution and the Related F-Distribution. A mathematical model which has played an important role in modern statistical analysis is the z-distribution developed by Fisher.

The quantity z is equal to one-half the difference of the natural logarithms of two independent estimates of the same population variance, or to the difference of the natural logarithms of the corresponding standard deviations. This distribution serves as the model against which tests of significance of experimental results attained in the analysis of variance and in multiple regression problems (to be discussed later) are compared.

Thus, suppose we have two samples of sizes, N_1 and N_2, each drawn at random from one of two populations of variates normally distributed with equal population variances σ^2.

Compute from the two samples

$$s_1^2 = \frac{\sum_{i=1}^{N_1} (X_i - \bar{X}_1)^2}{n_1} \qquad \text{and} \qquad s_2^2 = \frac{\sum_{j=1}^{N_2} (X_j - \bar{X}_2)^2}{n_2}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the respective means; s_1^2 and s_2^2 are the respective variance estimates; and $n_1 = N_1 - 1$, $n_2 = N_2 - 1$. Then

$$z = \tfrac{1}{2} \log_e \frac{s_1^2}{s_2^2} = \log_e \frac{s_1}{s_2} \tag{3.22}$$

z is distributed in the form

$$y = y_0 \frac{e^{n_1 z}}{(n_1 e^{2z} + n_2)^{\frac{1}{2}(n_1+n_2)}} \tag{3.23}$$

where y_0 may be taken such that the area of the curve is unity, and the experimental value, z_0, may be compared with the model to determine the probability that values of z equal to or greater than z_0 could be obtained by random sampling errors alone. The probability P will be given by the area under the curve to the right of the ordinate erected at z_0.

Fisher (Ref. 3) has computed tables giving values of z corresponding to different values of n_1 and n_2, and P, namely, the 5, 1, and 0.10 per cent points of the z-distribution. It should be pointed out that the table gives the values of z at which ordinates cut off "tails" of 5, 1, and .10 per cent of the total area of the curve for values of n_1 and n_2 chosen so that n_1 corresponds to the number of degrees of freedom associated with the larger of the two estimates of variance.

The z-distribution is unimodal and symmetrical if $n_1 = n_2$. For large values of n_1 and n_2 and also for moderate values when n_1 and n_2 are equal or nearly equal, the distribution of z becomes nearly normal about a mean of zero with a standard deviation, or standard error,

$$\sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

It is to be noted that z is a Studentized function; hence it is especially appropriate for small samples. The z-test may be regarded as an extension of the t-test to situations where more than two variants are under comparison. In fact, Fisher (Ref. 2) has shown that the normal curve, the χ^2-distribution, and Student's distribution are included as special cases of the two-parameter family of curves represented by the z-distribution. For instance, since $z = \log_e (t)$, the values for $n_1 = 1$ in the table of z are the logarithms of the values for $P = .05$ and $P = .01$ in the table of t (Ref. 3).

Tables of the variance ratio

$$F = e^{2z} = \frac{s_1^2}{s_2^2} \tag{3.24}$$

are available (see Table IV, Appendix) and are coming to be more commonly used than the table of z, since the troublesome logarithmic transformation is thereby avoided. Against this advantage perhaps is the advantage of greater accuracy in the use of the z-tables when interpolations are required. Tables for seven points of the F distribution are now available (Ref. 6).

The Binomial Distribution in Sampling Theory. We have previously described the binomial distribution and indicated that the normal distribution may be used as an approximation to it (see page 27). Since this distribution plays such a significant role in sampling theory, it should be considered somewhat more broadly.

A Sampling Experiment Leading to the Binomial Distribution. We begin by presenting the results of a simple sampling experiment consisting of the tossing of 10 coins 512 times.

The record of this experiment is available in Table 19. The observed values for the several probabilities of success, that is, the proportion of tails, X_p, are given in column 2. The calculations for the mean and standard deviation of the number of successes are given in columns 3 and 4. It is found that the mean $\bar{X} = 0.5$ and that the standard deviation $s = .162$. The corresponding theoretical values are 0.5 and .156, respectively.

TABLE 19

THE TEST OF GOODNESS OF FIT OF THE THEORETICAL BINOMIAL DISTRIBUTION FOR THE OBSERVED DISTRIBUTION OF SUCCESSES (THE PROPORTION OF TAILS) FROM 512 TOSSES OF 10 COINS AT A TIME

X_p	f_0	fX	fX^2	f_t	$f_0 - f_t$	$\frac{(f_0 - f_t)^2}{f_t}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1.0 0.9	2 5 } 7	2.0 4.5	2.00 4.05	0.5 5.0 } 5.5	1.5	0.4091 ......
0.8	15	12.0	9.60	22.5	− *7.5*	2.5000
0.7	68	47.6	33.32	60.0	8.0	1.0667
0.6	105	63.0	37.80	105.0	0.0	0.0000
0.5	134	67.0	33.50	126.0	8.0	0.5079
0.4	95	38.0	15.20	105.0	−*10.0*	0.9524
0.3	55	16.5	4.95	60.0	− *5.0*	0.4167
0.2	23	4.6	0.92	22.5	0.5	0.0111
0.1 0.0	8 2 } 10	0.8 0.0	0.08 0.00	5.0 0.5 } 5.5	4.5	3.6818
Total	512	256.0	141.42	512.00	$\chi_0^2 = 9.5457$	

d.f. = 8; .30 > P > 20

Sample values:

$$\bar{X} = \frac{256}{512} = 0.5$$

$$s = \sqrt{\frac{141.42}{512} - .25} = .162$$

Population values:

$$\mu = 0.5$$

$$\sigma = \sqrt{\frac{.5 \times .5}{10}} = .156$$

In column 5 the theoretical values f_t are given. Finally, we tested the agreement between the observed and theoretical values by means of the χ^2-test [column (7)]. We wish to test for goodness of fit and enter the χ^2-table (Table III, Appendix) with $\chi_0^2 = 9.5457$ and 8 degrees of freedom. It is found that $P > .20$. We conclude that the observed distribution may be regarded as in accordance with the binomial distribution law; that is, the discrepancies between the observed and theoretical frequencies may be attributable to random sampling fluctuations. The theoretical basis of the binomial distribution is given below.

The Binomial Expansion. Assume that we take N random samples each of size n and in each of which a specific number t possess a given character a and the remainder $n - t$ do not possess the character. Let $\frac{t}{n} = p$ and $1 - \frac{t}{n} = q$; then the frequencies of samples with $t = 0, 1, 2, \ldots, n$ are given by terms in the series $N(q + p)^n$; that is,

$$N\left[q^n + nq^{n-1}p + \frac{n(n-1)p^2q^{n-2}}{1 \cdot 2} + \cdots \frac{n!p^tq^{n-t}}{t!(n-t)!} + \cdots + p^n\right] \tag{3.25}$$

The terms in the expansion $(q + p)^n$ are relative frequencies in the frequency distribution of all possible different samples, classified by number of successes, say t, that may be drawn from the population according to the rules of simple random sampling. The distribution may be called the *sampling distribution* of the number of successes, $t = 0, 1, 2, \cdots, t \cdots, n$. It is more commonly known as the *binomial distribution,* since it results from the expansion of $(q + p)^n$.

The mean of the distribution is given by

$$\mu = np \tag{3.26}$$

and the variance

$$\sigma^2 = npq \tag{3.27}$$

If instead of the actual number of t's in each sample the proportion of t's, that is, $\frac{1}{n}$th of the number in each sample, is recorded, the mean proportion of t's would be

$$\mu = p \tag{3.28}$$

and the variance

$$\sigma^2 = \frac{pq}{n} \tag{3.29}$$

The standard deviation of the sampling distribution or the standard error provides a basis for judging the exceptionalness of any obtained sample, as illustrated in the following example.

EXAMPLE 2. *The Measurement of Exceptionalness.* Assume that in a random sample of 50 individuals, 16 have a character, say A. Is this exceptional? It is known that in the general population 20 per cent possess the character A.

Of a random sample of 50 individuals the exact proportion who would be expected to have character A is given by the sum of the terms of the expansion of the binomial $(\frac{4}{5} + \frac{1}{5})^{50}$ from the seventeenth term onward. This proportion equals .031. This method, though exact, is extremely laborious. For this reason it is advantageous to use an alternative method.

If $p = q = \frac{1}{2}$ and n is large, the area under the appropriate section of a normal curve gives a close approximation to the point distribution of the binomial. Departures from the given conditions result in less accurate approximations. For example, if either p or q is small and n is not large, the approximation could be rather crude. A practical procedure for determining the relative values of p and q for a given n if the normal curve may be expected to represent the binomial is the following:

The mean of the distribution should be, say, three standard deviations from the start. Thus we want

$$\begin{aligned} np &> 3\sqrt{npq} \\ n^2p^2 &> 9npq \\ np &> 9(1-p) \\ p(n+9) &> 9 \\ p &> \frac{9}{n+9} \end{aligned}$$

If, for example, $n = 50$, then

$$p > \tfrac{9}{59} > .15$$

In using the normal curve as an approximation, we proceed as follows in the problem worked out above by the binomial expansion. Calculate:

$$\frac{X - np}{\sqrt{npq}} = \frac{15.5 - 10}{2.83} = 1.95$$

According to the normal table,

$$P = .025$$

This value is compared with $P = .031$ above.

The Sampling Distribution of Differences between Percentages. Frequently, the experimental results relate to the case of two samples where it is desired to know whether the two samples may be regarded as random samples from the same population. Thus:

(1) In sample 1 of size n_1, there are t_1 individuals that have the character A.

(2) In sample 2 of size n_2, there are t_2 individuals that have the character A.

Could the two samples be random samples from populations in which p (the probability of character A occurring) is the same? Thus:

$$\frac{t_1}{n_1} = \frac{t_2}{n_2} = p$$

The theoretical or mathematical model against which such experimental results may be compared can be built up as follows: Assuming that $p_1 = p_2 = p$, the variation in t in repeated samples of n_1 follows the binomial $(q + p)^{n_1}$; similarly, the variation in t_2 in repeated samples of n_2 follows the binomial $(q + p)^{n_2}$.

t_1 varies about a mean of n_1p with a standard deviation, $\sqrt{n_1pq}$; t_1/n_1 varies about a mean of p with a standard deviation

$$\frac{1}{n_1}\sqrt{n_1pq} = \sqrt{\frac{pq}{n_1}}; \qquad E\left[\frac{t_1}{n_1} - p\right] = 0$$

That is, the mean in repeated sampling is p;

$$E\left[\frac{t_1}{n_1} - p\right]^2 = \frac{pq}{n_1}.$$

Similarly,

$$E\left[\frac{t_2}{n_2} - p\right] = 0; \qquad E\left[\frac{t_2}{n_2} - p\right]^2 = \frac{pq}{n_2}$$

Also,

$$E\left[\left(\frac{t_1}{n_1} - p\right)\left(\frac{t_2}{n_2} - p\right)\right] = 0$$

Consider:

$$d = \frac{t_1}{n_1} - \frac{t_2}{n_2}$$

$$\begin{aligned}[d^2] &= \left[\left(\frac{t_1}{n_1} - p\right)^2 - 2\left(\frac{t_1}{n_1} - p\right)\left(\frac{t_2}{n_2} - p\right) + \left(\frac{t_2}{n_2} - p\right)^2\right] \\ &= \frac{pq}{n_1} + \frac{pq}{n_2}\end{aligned}$$

$$\sigma_d^2 = [d^2] = pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{3.30}$$

$$\sigma_d = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{3.31}$$

The ratio

$$\frac{d}{\sigma_d} = \frac{\dfrac{t_1}{n_1} - \dfrac{t_2}{n_2}}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}. \tag{3.32}$$

in repeated sampling will be approximately normally distributed about 0 with unit standard deviation. The normal model may therefore be used for comparing the experimental results. The complete procedure is

1. Assume the hypothesis $p_1 = p_2 = p$.
2. Estimate p from the data; the maximum likelihood estimate is $p = \dfrac{t_1 + t_2}{n_1 + n_2}$.

3. Calculate the ratio $\dfrac{\dfrac{t_1}{n_1} - \dfrac{t_2}{n_2}}{\sqrt{pq\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$.
4. Refer to the normal probability scale; consider whether the results are compatible with the hypothesis.

PROBLEMS

The following problems are designed to give the student an understanding of sampling and sampling errors. The following normal population of numbers with a mean of 30 and a variance of 100 may be used for the exercises. Write or type each of the 100 numbers on a small square of cardboard or stiff paper. Place the 100 pieces in a box and mix thoroughly. Then draw one card at random from the box and record the number on it. Return the card to the box. Mix the cards again, draw a second card and record its number, and so on until a sample of a specified size has been obtained.

FREQUENCY DISTRIBUTION OF NUMBERS, X'S, IN A NORMAL POPULATION WITH $\mu = 30$; $\sigma^2 = 100$

X	f	X	f	X	f	X	f
57	1	39	3	28	3	15	1
53	1	38	2	27	3	14	1
49	1	37	2	26	4	13	1
48	1	36	3	25	3	12	1
47	1	35	3	24	3	11	1
46	1	34	3	23	2	7	1
45	1	33	5	22	2	3	1
44	1	32	2	21	3		
43	2	31	4	20	2		
42	3	30	10	19	3		
41	3	29	4	18	3		
40	2			17	2		
				16	1		

Total 100

Exercise: Selecting 20 samples of 10 at random,

1. Compute 20 means.
2. Compute 20 variances.
3. Combine to make 10 random sets of paired values of the means; of the variances.
4. Compute 10t's for differences between means of uncorrelated measures.
5. Take 10 samples of 5 in pairs and calculate the correlation coefficients.
6. Combine the results of the individual students in each case, form the frequency distribution of the statistic, and plot the histogram. Calcu-

late the mean and standard deviation of each distribution and compare with the population and expected values.

References

1. Fisher, R. A., "Applications of 'Student's' Distribution," *Metron*, Vol. 5 (1926), pp. 90–104.
2. ———, "On a Distribution Yielding Error Functions of Several Well-Known Statistics," *Proceedings of the International Mathematical Congress*, Toronto, 1924, pp. 805–813.
3. ———, *Statistical Methods for Research Workers*, 10th ed. Edinburgh: Oliver and Boyd, Ltd., 1946.
4. Hotelling, Harold, "The Selection of Variates for Use in Prediction with Some Comments on the General Problem of Nuisance Parameters," *Annals of Mathematical Statistics*, Vol. XI (1940), pp. 271–283.
5. Mahalanobis, P. C., *et al.*, "Tables of Random Samples from a Normal Population," *Sankhya*, Vol. 1 (1934), pp. 289–328.
6. Merrington, Maxine, and Thompson, Catherine M., "Tables of Percentage Points of the Inverted Beta (F) Distribution," *Biometrika*, Vol. XXXIII (1943), pp. 73–88.
7. "Student," "The Probable Error of a Mean," *Biometrika*, Vol. VI (1908), pp. 1–25.

CHAPTER IV

THE TESTING OF STATISTICAL HYPOTHESES

The Role of the Hypothesis in Scientific Investigations. In the well-developed empirical sciences, scientific procedure is primarily concerned in deriving predictions the validity of which is tested by the results of experiments. The modern development of science has been much facilitated by the practice of using hypotheses in planning and guiding scientific inquiries. Even a casual study of the investigations of eminent scientists reveals that they were guided in their work by some theory and that this theory guided their observations and experiments. Where the theory proved inadequate, it was modified. Occasionally it was abandoned completely, but then another was sought to plan the action. Significant experimentation requires the guidance of a hypothesis, and a successful experimenter does not collect observations unguided by theory to which the facts are related. Bacon maintained that if enough instances are gathered and tabulated correctly, the principle which explains them will simply emerge without any hypothesis about them having been formed. This contention has not been proved by the experimenter.

The working hypothesis plays an important part in statistical research. It serves as a guide in planning the investigation; in determining what data to collect; in classifying, ordering, and reducing them; and finally as the basis for formulating the judgments with respect to it.

Similar to the working plan of Newton, the scientist who did not formulate hypotheses needlessly ("*hypotheses non fingo*"), or to the metaphysical requirements of Ockham, the logician who considered it needless to recur to many entities when it was possible to get along with fewer ones ("*nunquam ponenda est pluralitas sine necessitate*"), the statistician's preferred method is to test the simplest hypothesis and to hold to a minimum number of new quantities or constructs. Thus the preferred hypothesis used by the statistician in the examination of his data is that the apparent variations and the estimates of presumed effects may be attributable to random sample errors or to fortuitous factors rather than to the action of new causes. This hypothesis can be tested by the application of the theory of errors. It will be recalled that the statistical models previously described were constructed on the basis of sampling errors. As long as experimental results conform to these models, the hypothesis of chance (or, more specifically in this case, sampling errors) being the cause of the observed effects is accepted.

The hypothesis that chance factors may have given rise to an observed effect is frequently spoken of as the *null hypothesis.* This hypothesis is met with in a number of different forms in research or statistical work. In experimentation, for example, it is often desirable to compare the effects of various methods of treatment or of production. The null hypothesis can in these cases be stated as follows: There is no difference in the outcomes of the several treatments, or, The outcomes are the same. This method is equivalent to determining whether or not the observed difference should be ascribed to random fluctuations or judged to be significant, that is, ascribed to the differential treatment. The null hypothesis assumes the former alternative, which, if found to be incompatible with the facts of observation, is then rejected. More generally, the null hypothesis may be stated thus: Can the samples under examination be regarded as having been randomly chosen from the same or similar population?

General Theory of Testing Statistical Hypothesis. In Chapter III, Sampling Distributions, it was stated that the statistician developed mathematical models against which the research worker could compare his experimental results and draw conclusions with respect to their significance. The process of determining statistical significance was said to consist in comparing the numerical data (or some function of them) obtained in a particular experiment with the model to establish whether or not they conform to the model. The name applied to the process of examining the significance of the data is the *test of significance.* In dealing with the sampling distribution we were concerned with testing the agreement between the distribution of our set of sample values and a theoretical distribution. In this case, we spoke of a test of the *goodness of fit.*

More recently, we have come to speak of the problem of testing statistical hypotheses and thus to speak of the test of significance relative to the hypothesis in question. Before proceeding to illustrate the application of these tests to some practical problems met with by the research worker, we shall describe briefly the theoretical basis underlying current procedures in testing a statistical hypothesis.

Suppose that a random variable X is the measurement of a certain character and that a number of repeated measurements are made, say N times. We thus obtain N random variables $X_1, X_2, \ldots, X_N$. The N random variables are assumed to be independently distributed, and the set of values is said to be a sample of N independent observations on X. The sample of N observations may be represented as a *sample point* E, in the N-dimensional space having as its coordinates $X_1, X_2, \ldots, X_N$. The space in which the point lies may be called the *sample space,* W.

Assume that the distribution of X is normal but that the values of some parameters $\theta_1, \ldots, \theta_q$ specifying the population are unknown. Any assumption about the unknown parameters $\theta_1, \ldots, \theta_q$ may be called a *statistical hypothesis.* The statistical hypothesis, H_1, is called a *simple hypothesis* if it determines completely the values of all the q-parameters, for example, if it specifies that $\theta_1 = 1$, $\theta_2 = 3$, $\cdots$. If the hypothesis is consistent with more values than one for some parameter it is called a *composite hypothesis;* for instance, the hypothesis that $\theta_1 = \theta_2$ for a distribution of X determined by two unknown parameters is a composite hypothesis.

For simplicity, we shall consider the case of a single unknown parameter. That is, let us assume that only one unknown parameter, θ, is involved in the distribution function of X and θ, or $F(X_1, X_2, \ldots, X_N, \theta)$. We wish to test the null hypothesis, $H_0: \theta = \theta_0$ against the only admissible alternative hypothesis, $H_1: \theta = \theta_1$. For example, we may test the significance of the deviation in the mean of a sample on the basis of a random sample of N independent observations $X_1, \ldots, X_N$ from a normal population X. Then H_0 is the hypothesis that X is normally distributed about the mean θ_0 with standard deviation, σ, and H_1 is the hypothesis that X is normally distributed about the mean θ_1 with standard deviation, σ.

The testing of the statistical hypothesis involves the choice of a *region, w,* called *critical* in the sample space W. It also involves the decision to reject the hypothesis if and only if the sample point E falls in w. Therefore, the test of the statistical hypothesis, H_0, consists in rejecting H_0, when the sample point, E, falls within a specified critical region, w_0, and in accepting H_0 (or at least not rejecting it) if the point falls without w_0. The fundamental problem is, therefore, the specification of the critical region, w_0.

The principle upon which the choice of the critical region depends was first advanced by Neyman and Pearson (Ref. 3). It is based on the control and minimizing of two kinds of error involved in testing the hypothesis, H_0: (1) the unjust rejection of the hypothesis, described as an error of the first kind, and (2) the failure to reject the hypothesis when, in fact, it is incorrect, that is, when some other hypothesis, H_1, is true, designated as an error of the second kind.

The probability of an error of the first kind determined by the hypothesis under test, say H_0, is called the *size* of the corresponding critical region, w_0, and is given by $P\{E \epsilon w_0 | H_0\}$, that is, the probability that E, as determined by the observational values will fall within the region, w_0, as determined by the hypothesis, H_0. This probability may be designated by α.

The probability of an error of the second kind is $P\{E \epsilon (W - w_0) | H_1\}$ where $(W - w_0)$ is the set of all sample points outside w_0. It may be specified as β. This probability is called the power of the test with respect to H_1.

Neyman and Pearson (Ref. 4), assuming that $P(X_1, \ldots, X_n | H_0)$

and $P(X_1, \ldots, X_n|H_1)$ are the probability laws of the X's as fixed by the hypothesis, H_0, being tested, and by H_1, a single alternative, have shown that the region w_0, established by the inequality

$$P(X_1, \cdots, X_n|H_1) \geq kP(X_1, \cdots, X_n|H_0) \tag{4.01}$$

when $k > 0$ is a constant selected such that

$$P\{E\epsilon w_0|H_0\} = \alpha \tag{4.02}$$

is the best critical region with regard to H_1 having size α. The critical region, which provides the most powerful test with respect to H_1, is called the *best critical region* for H_0 with respect to H_1. ✓

This theory of testing statistical hypotheses is based on the simple principle of arranging the test, that is, of choosing the critical region, w_0, so as to minimize the probability of errors of the second kind while keeping the probability of errors of the first kind constant. The size of the critical region is then determined by α and its power is designated as $1 - \beta$. It is obviously impossible to make both α and β arbitrarily small. The decision of just how the balance between the two kinds of errors should be struck must be made by the investigator and will presumably be based on the relative importance of the two kinds of error in the particular situation. It is the function of statistical theory to show how the two risks of error may be controlled and minimized.

In practice, the investigator controls the first kind of error by choosing as the value of α, the boundary of the critical region, a specified level of significance, say the 5 per cent, 1 per cent, or 0.1 per cent point value of the criterion. The level is decided upon at the time of designing the investigation and depends on the nature of the problem and the risk in error the investigator is willing to accept. The custom is to reject the hypothesis tested if the observed value of the criterion is greater than (lies beyond, usually) the 1 per cent point, to remain in doubt if it lies between the 5 per cent and 1 per cent points, and to accept the hypothesis if the criterion is less than the 5 per cent point. With respect to the control of the second type of error, studies of the power function of tests have been made and tables are available for securing the probability of errors of the second kind in some instances. Neyman and Tokarska (Ref. 5) have compiled tables for use in determining the probability of errors of the second kind in testing Student's hypotheses. Tang (Ref. 6) has tabled the power function for the test of general linear hypotheses, which reduces to Fisher's z-test. Lehmer (Ref. 2) has prepared further tables for detecting the probability of errors of the second kind in dealing with linear hypotheses. Eisenhart (Ref. 1) investigated the power function of the χ^2-test.

The relation between the probabilities of the two kinds of error involved in testing the hypothesis, $H_0: \theta = \theta_0$, against the alternative, $H_1: \theta = \theta_1$, is illustrated in Fig. 4.

The probability of accepting the hypothesis, $H_0:\theta = \theta_0$ when it is true, is given by $1 - \alpha$. That is, the critical region, w_0, is the area to the right of the ordinate erected at $X = X_0$ in the θ_0-curve; the probability of accepting the hypothesis, $H_1:\theta = \theta_1$ when it is true, is given by β, the area under the θ_1-curve which lies to the right of the ordinate at $X = X_0$. The quantity β relative to θ_0, θ_1, and α as defined previously is the power of the test which specifies w_0 as the critical region. Hence, α and (1-β) represent the probabilities of the first and second kinds of error, respectively.

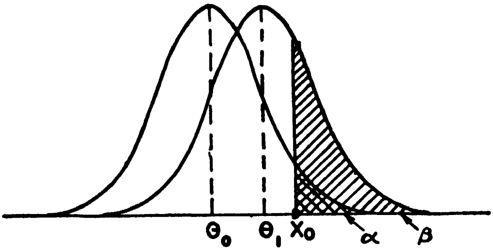


Figure 4. Normal distributions of the univariates $p(x,\theta_1)$ and $p(x,\theta_0)$ with critical regions for testing alternative hypotheses relative to the mean.

Neyman and Pearson use a criterion based on the principle of likelihood as the basis for accepting or rejecting a given hypothesis. In the case of the hypothesis tested above, H_0, the ratio

$$\lambda = \frac{P_0(X_1, X_2, \cdots, X_n)}{P_1(X_1, X_2, \cdots, X_n)} \tag{4.03}$$

is designated as the likelihood of the hypothesis, H_0, as tested against the single alternative hypothesis, H_1.

In accordance with Equation (4.03), a most powerful region, α, is comprised of all points which satisfy the inequality

$$\frac{P(X_1, \cdots, X_n|H_1)}{P(X_1, \cdots, X_n|H_0)} \geq k \tag{4.04}$$

where k is selected so that the region should have the required size α as indicated in Equation (4.02). For example, the principle for choosing the critical region, w_0, may be applied to the case of testing the significance of a mean of a sample from a normal population, where $H_0:\theta = \theta_0$ and $H_1:\theta = \theta_1$. We specify that the critical region required and defined by the inequality [Equation (4.04)] has the size $\alpha = .01$.

Since, under the hypothesis H_0, the variate

$$\frac{1}{N} \sum_a (X_a - \theta_0) \quad (a = 1, \cdots, k) \tag{4.05}$$

is normally distributed about a mean of zero with variance $1/N$, k can be read from a normal table:

$$k = \frac{2.326}{\sqrt{N}} \tag{4.06}$$

The most powerful region of size .01 is then

$$\sum_{\alpha} (X_\alpha - \theta_0) \geq \frac{2.326}{\sqrt{N}} \tag{4.07}$$

that is, the test specified by the region (4.07) is most powerful with regard to all alternatives, $\theta > \theta_0$.

If the probability of an error of the first kind, α, and of the second kind, β, is specified in a given problem, it is possible to determine the minimum size of sample, N, for which the power of the most powerful region of size α is equal to or greater than $1 - \beta$. For testing H_0 against H_1, for instance, the minimum number of observations is equal to the smallest positive integer, N, for which

$$\beta_N(\alpha) \leq \beta \tag{4.08}$$

where $\beta_N(\alpha)$ denotes that for a fixed N, β is a single-valued function of α. For example, (1) if the arithmetic mean, $\bar{X}$, of a predetermined number of N observations is less than or equal to a properly selected constant, k, the hypothesis being tested, H_0, is accepted; and (2) if $\bar{X} > k$, the hypothesis, H_0, is rejected. N and k are determined such that the probability of (1) is equal to $1 - \alpha$ when $\theta = \theta_0$ and is equal to β when $\theta = \theta_1$.

Sequential Test of a Statistical Hypothesis. Recently a test has been developed whereby the number of observations is not predetermined but is kept as a random variable. Instead of deciding in advance the number of items to be included in a sample, the data are analyzed continuously as they are being collected (Ref. 7). In such cases where it is possible to examine the data as they originate, as in some manufactured products, the *sequential probability-ratio test* frequently uses half as many observations as the current most powerful test. Briefly, the principal properties of the sequential test are as follows:

(1) The procedure by which a sequential test of a statistical hypothesis is carried out depends on the following rule of behavior:

(a) To accept the null hypothesis being tested.
(b) To reject the hypothesis.
(c) To suspend judgment, that is, to continue the analysis by making an additional observation.

The test procedure is kept up sequentially until either decision (a) or (b) is made.

(2) If α is the probability that when H_0 is true, the alternative hypothesis, H_1, will erroneously be accepted, and if β is the probability that when H_1 is true, H_0 will falsely be accepted, then it is necessary that

$\alpha + \beta < 1$. Sequential analysis determines in the course of the analysis whether or not the data justify a decision with a risk in error of judgment as small as α or β. The number of observations necessary will, on the average, depend on how small α and β are made; also on how fine a distinction is made between H_0 and H_1.

(3) The fundamental criterion basic to the decision in (1) is the likelihood ratio, L, which is the ratio of the probability that the one hypothesis truthfully specifies the origin of the observed data to the probability that the alternative hypothesis does. The value of L required to accept H_0 is $\frac{1 - \beta}{\alpha}$; that required to accept H_1 is $\frac{\beta}{1 - \alpha}$. L is computed after each observation and is compared with the critical values necessary for a decision. These values of L are independent of the number of observations. Since the likelihood ratio, as used in sequential tests, is a continuing product, considerable saving in calculation results by using log L instead of L. .

In practice, the quantities α and β are usually taken as quite small, rarely greater than .05 and frequently .01 or less.

References

1. Eisenhart, Churchill, "The Power Function of the χ^2-test," *Bulletin of the American Mathematical Society*, Vol. 44 (1938), p. 32 (abstract).
2. Lehmer, Emma, "Inverse Tables of Probabilities of Errors of the Second Kind," *Annals of Mathematical Statistics*, Vol. XV (1944), pp. 388–398.
3. Neyman, J., and Pearson, E. S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society (London)*, Series A, Vol. CCXXXI (1933), pp. 289–337.
4. ———, and ———, "Contributions to the Theory of Testing Statistical Hypotheses, *Statistical Research Memoirs*, Vol. I (1936), pp. 1–37.
5. ———, and Tokarska, B., "Errors of the Second Kind in Testing 'Student's' Hypothesis," *Journal of the American Statistical Association*, Vol. 31 (1936), pp. 318–326.
6. Tang, P. C., "The Power Function of the Analysis of Variance Tests with Tables and Illustrations of Their Use," *Statistical Research Memoirs*, Vol. II (1938), pp. 126–157.
7. Wald, A., "Sequential Tests of Statistical Hypotheses," *The Annals of Mathematical Statistics*, Vol. XVI (1945), pp. 117–186.

CHAPTER V

CURRENT PROCEDURES IN TESTING STATISTICAL HYPOTHESES

Up to this point, we have defined a number of statistical models against which the research worker may compare his experimental results. We have also discussed the theoretical formulation and solution of the problem of testing statistical hypotheses. It is now the purpose to show how, in a given situation when faced with some practical problem, the research worker may utilize the principles underlying the theory in deciding which model, if any, is applicable in his particular problem, and how to choose it intelligently and effectively. This chapter will be devoted to illustrating ways of solving a number of problems most of which are of frequent occurrence.

Problem V.1. The significance of a mean from a known normal population. The simplest case of testing significance is in the problem where the population is known, that is, the population parameters, the mean and standard deviation, are known and the quantity whose significance we are interested in testing may be assumed to be normally distributed in the population. Specifically, the question is: Could this sample be a random sample from that population? Such, for instance, is the problem of determining whether or not a given sample of pupils to whom an intelligence test has been given could be regarded as a random sample from the population upon whom the norms of the test were set by the author.

Assume that it is known that for a particular intelligence test the I.Q.'s are normally distributed about a mean of 100 with a standard deviation of 17 I.Q. points. The test is administered to a class of 36 pupils who in other respects appeared to belong to this population. The mean I.Q. for the class was found to be 108. May we conclude that the class is a random sample from the specified population—that the mean ability of the class is the same as that of the population?

To answer this question we need to determine what model should be used with which the experimental result can be compared. It is known from sampling theory (page 36) that the means of samples of 36 cases drawn at random from this population will be normally distributed about the population mean, 100 I.Q. points, with a standard deviation (or standard error) equal to $\sigma/\sqrt{N} = \frac{17}{6} = 2\frac{5}{6}$ I.Q. points. We found a mean of 108 I.Q. points. How often should we expect to find a mean as high as this or higher in repeated sampling from this population?

The answer is obtained by referring to the normal probability table (Table I, Appendix). To enter this table we must convert the raw score to a standard measure. Thus:

$$z = \frac{108 - 100}{2.833} = 2.82$$

From the table, we find that in repeated sampling from this population we should expect to find a value as high as or higher than the one obtained in 1 − .9976, or 0.24 per cent, of cases. This probability is lower than the level of 1 per cent which we decided to use. Therefore, we conclude that the sample could not have been drawn from the specified population. We are aware that in making this statement we shall be wrong in 0.24 per cent of the cases; but this is a risk we are willing to run. Such is the statistical conclusion. The education conclusion is that the mean ability of the class tested is significantly above the norm specified for the population.

Problem V.2. The significance of a mean from an unknown normal population. We shall take next the problem in which the population mean is known, or specified by hypothesis to be some value, say 100 I.Q. points, but in which the population standard deviation is not known.

We gave an intelligence test to a class in Grade 5 consisting of 26 pupils. The mean I.Q.-score on the test was 93. We want to know if our class may be assumed to be a random sample from a population whose mean, μ, equals 100 I.Q. points. To answer this question we need to compare our result with the appropriate model, which in this case must be the distribution of t, (see page 43), since we do not know the population standard deviation. Therefore, we calculate the value of t, say t_0, for our sample and compare it with the t-model. If we find that the probability of getting a value of t greater than or equal to $\pm t_0$ in repeated sampling is less than 1 in 100, then we conclude that the sample could not have been drawn from this population.

The necessary calculations and procedures are as follows:

$$N = 26; \qquad \bar{X} = \frac{\Sigma X}{N} = 93; \qquad s^2 = \frac{\Sigma(X - \bar{X})^2}{(N - 1)} = 144$$

The value of t_0 is

$$t_0 = \frac{(\bar{X} - \mu)}{\sqrt{\frac{s^2}{N}}} = \frac{-7}{2.353} = -2.97$$

We compare this value of t_0 with the model as given in the table of the t-distribution (Table II, Appendix). We enter the row of the table corresponding to $n = N - 1$, that is, $n = 25$ in our example. For samples of 26 ($n = 25$), we expect to find values of t greater than or equal to ± 2.787 in 1 per cent of cases; so, clearly, we should expect to find values greater than or equal to $\pm t_0 = \pm 2.97$ in repeated sampling from this population in an even smaller percentage of cases. Our conclusion, therefore, is that it is unlikely that our sample was drawn from a population in which the mean I.Q. was 100, or, that the mean ability of the class is significantly different from the norm of 100.

Problem V.3. The significance of a mean from a small finite population. In most sampling problems a large population exists or is assumed to exist. At times the problem arises of using a sample which may comprise an appreciable part of a relatively small finite population sampled. The standard error of the mean is then adjusted as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N - n}{N - 1}} \tag{5.01}$$

This adjustment follows from the fact that sampling errors affect only the estimate of that fraction of the whole which is not included in the sample. The value of σ, the population standard deviation, is usually unknown, but the unbiased estimate of it can be obtained from the sample. N is the size of the population; n, the number of sampling units.

For example, suppose a sample of 50 female students has been drawn at random from the 500 female freshmen enrolled in a university. We wish to test the hypothesis that the mean height of the 500 freshmen is equal to 168 cm.

We calculated the following statistics for the sample of 50:

$$\begin{aligned} \bar{X} &= 164.8 \text{ cm} \\ s_X &= 5.9 \text{ cm} \end{aligned}$$

Then

$$\begin{aligned} s_{\bar{x}} &= \frac{5.9}{\sqrt{50}}\sqrt{\frac{500 - 50}{499}} \\ &= .79 \\ t_0 &= \frac{\bar{x} - \mu}{s_{\bar{x}}} \\ &= \frac{164.8 - 168}{.79} = -4.05 \end{aligned}$$

We compare the value t_0 with the t-model. Entering the table of the t-distribution (Table II, Appendix) with $n = N - 1$, or 49, we find that for $n = 40$, $-t_{.0005} = -3.551$ and that for $n = 60$, $-t_{.0005} = -3.460$. Since our value is obviously greater than the tabled values, we may reject the hypothesis that the mean height of the 500 freshmen is equal to 168 cm. If the statement that $\mu = 168$ were true, we should expect that in repeated sampling, 50 students selected at random from the 500 would give a mean as divergent as 164.8 less than once in 2000 trials.

Problem V.4. The significance of the difference between means. More frequently the problem is that of testing whether or not there is a significant difference between means, that is, whether or not the samples may be regarded as random samples from the same normal population; to test the hypothesis that the true difference between means is zero.

The experimental results may also in this case be compared with the model t-distribution in making the test of significance (see page 47), because (1) the difference between two means may be regarded as normally distributed about zero (if the hypothesis is true) with a standard deviation σ, and (2) the standard error of the difference estimated on the number of degrees of freedom provides an independent estimate of σ. Since t, in general, is the ratio of (1) to (2), it is the appropriate criterion for the test of the hypothesis involved here.

The following calculations and procedures enable us to make the determination of the t_0 in this particular case (the subscripts refer to the corresponding sample):

$$\bar{X}_1 = \frac{\Sigma X_1}{N_1}; \qquad \bar{X}_2 = \frac{\Sigma X_2}{N_2}$$

$$s^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{N_1 + N_2 - 2}$$

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{s^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

Refer t_0 to the table of t (Table II, Appendix).

Let us illustrate by taking the problem to test if two sets A and B of test scores from two classes in algebra may be regarded to have come from the same normal population. We obtain the following values:

Class A	*Class B*
$N_1 = 34$	$N_2 = 30$
$\Sigma X_1 = 975$	$\Sigma X_2 = 795$
$\bar{X}_1 = 28.68$	$\bar{X}_2 = 26.50$
$\Sigma(X_1 - \bar{X}_1)^2 = 4327.4$	$\Sigma(X_2 - \bar{X}_2)^2 = 2969.5$

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{N_1 + N_2 - 2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$$= \frac{28.68 - 26.50}{\sqrt{\frac{4327.4 + 2969.5}{34 + 30 - 2}\left(\frac{1}{34} + \frac{1}{30}\right)}}$$

$$= \frac{2.18}{\sqrt{117.69 \times .062745}} = .802$$

We enter the t-table in the row corresponding to $n = N_1 + N_2 - 2$, to find the probability of obtaining a value of t greater than or equal to $\pm t_0$ in repeated sampling. In the example, $n = 34 + 30 - 2 = 62$, but this specific value is not given in the table. It is observed from the values for $n = 60$ and $n = 120$ that the probability of getting a value of t greater than or equal to ± 0.802 in repeated sampling is somewhere between .40 and .50. We conclude, therefore, that the two classes may be assumed to be random samples from the same normal population or, in other words, that the means of the two classes are not significantly different. The pedagogical conclusion is that there is no real difference between the average algebraic abilities of the two classes as measured by the test used.

The hypothesis tested above, that the two samples were random samples from the same normal population, is equivalent to testing the hypothesis, H_1, that $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$, against the set of all the alternative hypotheses which specify only that $\mu_1 \neq \mu_2$ or $\sigma_1^2 \neq \sigma_2^2$, or both. General results have been obtained by Sato (Ref. 13, page 1) to indicate that the t-test is also the uniformly most powerful of all the unbiased exact tests that can possibly be made for the hypothesis, H_2: In connection with two uncorrelated normal populations, π_1 and π_2, it is assumed as given that σ_1 and σ_2 have the same (though unknown) value, to test the hypothesis, that $\mu_1 = \mu_2$, against the set of alternatives that $\mu_1 \neq \mu_2$. The study of the power function of the t-test under different conditions has been made by Hsu (Ref. 13).

We shall next consider a practical problem which occasionally arises when there is evidence to indicate that the variances of the two populations from which random samples have been drawn are unequal and it is desired to test the significance of the difference between the means.

Problem V.5. The significance of the difference between means when the variances are unequal or unknown. For a precise test of significance first given by Behrens (Ref. 3) for the difference between the means of two samples supposedly not drawn from equally variable populations, or from populations having a known variance ratio, the Behrens-Fisher method is available (Refs. 7, 8, 9, 22). Its application is made in the following example.

At the end of a certain course in science, two groups, one in U High School and one in B High School, took the Peterson Comprehensive Science Examination (Ref. 18). The following results were recorded:

School U: $N_1 = 14$ $\bar{X}_1 = 73.21,$ $(\text{S.D.})_1 = 21.53$
or $\Sigma(X_1 - \bar{X}_1)^2 = 6489.5726$
School B: $N_2 = 12$ $\bar{X}_2 = 56.30,$ $(\text{S.D.})_2 = 16.75$
or $\Sigma(X_2 - \bar{X}_2)^2 = 3366.7500$

From these data we obtain:

(a) $F = 1.6314,$ $n_1 = 13,$ $n_2 = 11:$ $.30 > P > .20$
(not significant)

(b)
$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\Sigma(X_1 - \bar{X}_1)^2}{N_1(N_1 - 1)} + \dfrac{\Sigma(X_2 - \bar{X}_2)^2}{N_2(N_2 - 1)}}} = 2.162 \qquad (5.02)$$

(c) For

$$n_1 = 13, \qquad n_2 = 11, \qquad \text{mean-variance ratio} = 1.6314 \times \tfrac{12}{14}$$
$$\theta = \tan^{-1}\sqrt{1.6314 \times \tfrac{12}{14}} = \tan^{-1}(1.1825) = 49°47'$$

We enter Sukhatme's table of the d-function (Ref. 22). We have $n_1 = 13$, $n_2 = 11$, $\theta = 49°47'$. No $d_{.05}$ (or $d_{.01}$) is given for these values. So we must find $d_{.05}$ to fit these values. We may interpolate for either n_1 or n_2 first; the result will be the same result in either case. Here we shall interpolate for n_1 first. For $n_1 = 13$ we get the following $d_{.05}$ values:

θ	0°	15°	30°	45°	60°	75°	90°
$n_2 = 8$	2.306	2.293	2.261	2.225	2.196	2.176	2.170
$n_2 = 12$	2.179	2.175	2.167	2.163	2.164	2.168	2.170

Now we interpolate for $n_2 = 11$ and get the following $d_{.05}$ values:

$n_1 = 13, \quad n_2 = 11$

θ	0°	15°	30°	45°	60°	75°	90°
d_i	2.190	2.185	2.175	2.169	2.167	2.169	2.170

For $n_1 = 13$, $n_2 = 11$, $\theta = 45°$: $d_{.05} = 2.169$
For $n_1 = 13$, $n_2 = 11$, $\theta = 60°$: $d_{.05} = 2.167$

Since our observed $d_0 = 2.162$ is less than either of these $d_{.05}$ values and since our value of θ, 49°47′ is between 45° and 60°, there is no need for interpolating for θ. We now may declare our observed value of d nonsignificant at the 5 per cent level.

It is worth noting that if we had used the usual t-test for the hypothesis of equal means, the hypothesis would have been rejected at the 5 per cent level. Thus:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{N_1 + N_2 - 2}\left(\dfrac{N_1 + N_2}{N_1N_2}\right)}} = 2.121 \qquad (5.03)$$

For $n = N_1 + N_2 - 2 = 24$, $\quad P < .05$

An approximate method for the same problem was proposed by Cochran and Cox (Ref. 21), a method to test the hypothesis of equality of means with no hypothesis about the population variance when $N_1 \neq N_2$ and $s_1 \neq s_2$. In this test the variance of each mean is calculated separately. A criterion t is obtained by computing a weighted mean of the two t-values for the two samples, the weights being the two variances of the respective means. The ratio $\dfrac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}}$ is then compared with the weighted t-value to judge the significance.

The approximate test has been applied to the same data analyzed above by the Behrens-Fisher formula. The calculations are set forth in Table 20.

TABLE 20

Calculations for the Cochran-Cox Method of Testing the Significance of the Hypothesis of Equality of Means with No Hypothesis about the Population Variance

N	D.F.	$\bar{X}$	Σx^2	S^2	$S_{\bar{X}}^2$	$t_{.05}$	$t_{.05}S_{\bar{X}}^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
14	13	73.21	6489.5726	499.1959	35.6520	2.160	77.0191
12	11	56.30	3366.7500	306.0682	25.5057	2.201	56.1380
26	24	$D = 16.91$	9856.3226		61.1577		133.1571

$$\text{The criterion (weighted } t) = \frac{t_{.05}s_{\bar{X}_1}{}^2 + t_{.05}s_{\bar{X}_2}{}^2}{s_{\bar{X}_1}{}^2 + s_{\bar{X}_2}{}^2}$$

$$= \frac{77.0191 + 56.1380}{35.6520 + 25.5057}$$

$$= 2.177$$

The observed t is calculated as follows:

$$\frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{73.21 - 56.30}{\sqrt{35.6570 + 25.5057}} = 2.162$$

Since the observed t is less than the criterion t, that is, since $2.162 < 2.177$, the hypothesis of equal means is not rejected. Thus, the same conclusion is found as in the case of the exact test provided by the Behrens-Fisher formula.

Where the sizes of the samples are the same, that is, where $N_1 = N_2$, the significance of the difference between the means can be determined, even though the variances differ, by calculating the value of t in the usual way applying Formula (5.03). However, the t-table is entered with d.f. $= N_1 - 1 (= N_2 - 1)$ instead of $N_1 + N_2 - 2$.

Problem V.6. The significance of the difference between the means of correlated measures. Situations arise in which the two samples are equal in number and in which each individual of one sample corresponds in some way to a particular individual of the second sample. Such is the case, for example, when individuals have been paired or equated on certain characteristics, in two different groups. One group is then subjected to one type of treatment and the second to another. At the end of the experimental period, evidence is obtained as to whether a differential effect has resulted. In this case and in others of a like kind, we can use the distribution of t as the theoretical model. It is necessary, however, to calculate t_0 in a way different from that illustrated in Problem

V.4. As before, we wish to determine whether the two groups may be regarded as random samples from the same normal population or to test the null hypothesis that the two means are the same. The individuals in the two groups have been equated, a fact that must be taken into account when setting up the model. If there is no differential effect, then clearly the difference between the criterion measures of the paired individuals should be zero.

In practice, as has been noted, in taking means of samples from the same population, the differences will never be exactly zero, even if there is no differential effect. The distribution of differences between corresponding values now constitutes a single sample, for which the mean difference and the standard error of the mean difference can be calculated in the usual manner. The ratio of the mean difference to its standard error will be distributed as t in repeated sampling. Therefore, the distribution of t is the theoretical model against which to check the experimental results.

The following data were obtained in an experiment to compare the efficacy of two methods of teaching elementary algebra to high-school classes. One group was taught by the group method, the other by the individual method. The individuals constituting each of the 25 pairs were equated on the basis of intelligence test scores and mathematical pretests.

Our problem is to test the null hypothesis that there is no difference between the two teaching methods with respect to the outcomes measured. This is equivalent to determining from the experimental data whether the mean scores on the criterion of the two groups are the same, that is, whether the two classes may be assumed to be random samples from the same normal population. If it is found that the mean scores are significantly different, the conclusion will be drawn that there is evidence of a differential effect between the two methods of teaching.[1]

The data are recorded in Table 21. We first calculate the differences between the scores made by the paired individuals. These differences are given in column (4). We then find the mean of the distribution of differences:

Mean differences,

$$\bar{D} = \bar{X}_D = \frac{\Sigma D}{N} = \frac{232}{25} = 9.28$$

We next calculate the variance of the differences:

$$\begin{aligned} s_D^2 &= \frac{N\Sigma D^2 - (\Sigma D)^2}{N(N-1)} \\ &= \frac{25 \times 8962 - (232)^2}{(25)(24)} \\ &= 283.71 \end{aligned}$$

[1] For a rigorous discussion of the single-factor experiment, see page 286.

TABLE 21

CALCULATIONS FOR TESTS OF SIGNIFICANCE OF DIFFERENCE IN PAIRED GROUPS BY TWO METHODS

Pair No.	Achievement score		Difference			(2) − 50		(3) − 50					
	Experimental	Control	D		D^2	X		Y		XY		X^2	Y^2
(1)	(2)	(3)	(4)		(5)	(6)		(7)		(8)		(9)	(10)
			−	+		−	+	−	+	−	+		
I	73	58		15	225		23		8		184	529	64
II	52	37		15	225		2	13		26		4	169
III	100	53		47	2,209		50		3		150	2,500	9
IV	60	77	17		289		10		27		270	100	729
V	75	51		24	176		25		1		25	625	1
VI	67	62		5	25		17		12		204	289	144
VII	61	55		6	36		11		5		55	121	25
VIII	59	30		29	841		9	20		180		81	400
IX	33	39	6		36	17		11			187	289	121
X	19	16		3	9	31		34			1,054	961	1,156
XI	32	15		17	289	18		35			630	324	1,225
XII	27	37	10		100	23		13			299	529	169
XIII	68	44		24	576		18	6		108		324	36
XIV	54	27		27	729		4	23		92		16	529
XV	26	43	17		289	24		7			168	576	49
XVI	30	27		3	9	20		23			460	400	529
XVII	69	53		16	256		19		3		57	361	9
XVIII	43	29		14	196	7		21			147	49	441
XIX	23	13		10	100	27		37			999	729	1,369
XX	11	17	6		36	39		33			1,287	1,521	1,089
XXI	26	20		6	36	24		30			720	576	900
XXII	30	9		21	441	20		41			820	400	1,681
XXIII	28	35	7		49	22		15			330	484	225
XXIV	53	21		32	1,024		3	29		87		9	841
XXV	23	42	19		361	27		8			216	729	64
Total	1,142	910	82	314	8,962	299	191	399	59	493	8,262	12,526	11,974
			232			−108		−340		7,769			

$$\bar{X} = \frac{1142}{25} = 45.68 \qquad \text{Check: } \bar{X} = \frac{-108}{25} + 50 = 45.68$$

$$\bar{Y} = \frac{910}{25} = 36.40 \qquad \text{Check: } \bar{Y} = \frac{-340}{25} + 50 = 36.40$$

$$\bar{D} = \frac{232}{25} = 9.28 \qquad \text{Check: } \bar{X} - \bar{Y} = 45.68 - 36.40 = 9.28$$

TABLE 21 (*Continued*)

Method 1

$$\begin{aligned}
\sigma_{\substack{\text{mean}\\ \text{diff.}}} &= \sigma_{\bar{D}} = \sqrt{\frac{N\Sigma D^2 - (\Sigma D)^2}{N^2(N-1)}} \\
&= \sqrt{\frac{25 \times 8962 - (232)^2}{25^2.24}} \\
&= 3.368 \\
t_0 &= \frac{9.28}{3.37} \\
&= 2.754 \\
n &= N - 1 = 24 \\
P &< .05 \\
\text{or } .02 &> P > .01 \\
t_{.02} &= 2.492;\ t_{.01} = 2.797
\end{aligned}$$

Method 2

$$\begin{aligned}
\sigma_{\bar{X}-\bar{Y}} &= \sqrt{\frac{\sigma_X{}^2}{N} + \frac{\sigma_Y{}^2}{N} - 2r_{\bar{X}\bar{Y}}\frac{\sigma_X\sigma_Y}{N}} \\
&= \sqrt{\frac{N\Sigma x^2 + N\Sigma y^2 - 2N\Sigma xy}{N^2(N-1)}} \\
N\Sigma x^2 &= N\Sigma X^2 - (\Sigma X)^2 = (25)(12{,}526) \\
&\qquad - (-108)^2 = 301{,}486 \\
N\Sigma y^2 &= N\Sigma Y^2 - (\Sigma Y)^2 = (25)(11{,}974) \\
&\qquad - (-340)^2 = 183{,}750 \\
2N\Sigma xy &= 2(N)(\Sigma XY) - (\Sigma X)(\Sigma Y) \\
&= 2(25)(7769) - (-108)(-340) \\
&\qquad = 2(157{,}505) \\
&= 315{,}010 \\
N^2(N-1) &= (25)^2\,(24) = 15{,}000 \\
\sigma_{\bar{X}-\bar{Y}} &= \sqrt{\frac{301{,}486 + 183{,}750 - 315{,}010}{15{,}000}} \\
&= \sqrt{11.3484} = 3.368 \\
t_0 &= \frac{9.28}{3.37} \\
&= 2.754 \\
n &= N - 1 = 24 \\
P &< .05;\ .02 > P > .01
\end{aligned}$$

The variance of the mean is then

$$\begin{aligned}
s_{\bar{D}}{}^2 &= \frac{s_D^2}{N} \\
&= \frac{283.71}{25} = 11.348
\end{aligned}$$

The standard error of the mean is

$$s_{\bar{D}} = \sqrt{11.348} = 3.37$$

In one operation, the calculation of $s_{\bar{D}}$ is

$$\left.\begin{aligned}
s_{\bar{D}} &= \sqrt{\frac{N\Sigma D^2 - (\Sigma D)^2}{N^2(N-1)}} \\
&= \sqrt{\frac{224{,}050 - 53{,}824}{(625)(24)}} = 3.37
\end{aligned}\right] \qquad (5.04)$$

Then

$$t_0 = \frac{\bar{D}}{\sqrt{\dfrac{\Sigma(D - \bar{D})^2}{N(N-1)}}} = \frac{9.28}{3.37} = 2.7537 \qquad (5.05)$$

From the table of t, entering the row corresponding to $n = N - 1 = 24$, we find the chance of getting a value of t greater than or equal to $\pm t_0$; that is, ± 2.754 is slightly greater than 1 in 100 ($t_{.01} = 2.797$). Hence, the null hypothesis is rejected at the 5 per cent level of significance. We conclude that the two groups can not be assumed to be random samples from the same normal population, or that the mean scores are significantly different at the 5 per cent level. The educational conclusion under certain assumptions is that the two methods of teaching produced significantly different results.

When a large number of pairs of individuals are used it may be advantageous to work with the original measures, thus avoiding the calculation of the differences.

Using this method of calculation, the value of t_0 is

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2 - 2\Sigma(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{N(N-1)}}} \qquad (5.06)$$

The calculations are shown for the same problem used to illustrate the method based on the calculation of the differences. This method follows (with appropriate methods for reducing the mathematical calculations) the more commonly used formula of the standard error of the difference between the means of correlated measures:

$$\sigma_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_x^2 + \sigma_y^2 - 2r_{\bar{x}\bar{y}}\sigma_x\sigma_y}{N}} \quad \begin{pmatrix}\text{when } X = X_1 \\ \text{and } \ Y = X_2\end{pmatrix} \qquad (5.07)$$

The demonstration of the equivalence by applying the respective methods to the same set of data is given in Table 21. It should be noted that the unbiased estimate of σ^2 is used in both cases.

If we had not utilized the information provided by the experimental design, different results would have been obtained as noted below. Using the method for testing the significance of the difference between the means of random samples as in Problem V.4, we have, since $N_1 = N_2$,

$$\left.\begin{aligned} t_0' &= \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{N(N-1)}}} \\ &= \frac{9.28}{\sqrt{32.3491}} = 1.6 \end{aligned}\right] \qquad (5.08)$$

Entering the table of t with $n = 2(N - 1) = 48$, it is observed (without interpolation) that we should expect to get a value of t greater than or equal to $\pm t_0'$, i.e., ± 1.6 in repeated sampling in more than 5 per cent of the cases. The conclusions are, therefore, altered from those drawn earlier by the calculation of t_0.

It is usually advisable to calculate both t_0 and t_0' and make both tests of significance. Sometimes one and at other times the other may be the more sensitive. If either one or the other shows a significant difference between the means, it is safe to accept the conclusion of significance. If as is most often the case in experimental work, the corresponding values for the respective paired individuals are positively correlated, the standard deviation of the differences will be thereby reduced. Against this favoring circumstance must be weighed the fact that in treating the results as a single sample, the number of degrees of freedom is only half as great as if the two samples had been treated separately, that is, if two random samples were used for experimental subjects. From the findings of the two tests of significance for the same data, a direct statistical measure of the efficacy of the basis of pairing used is made available.[2]

Problem V.7. The sign test of significance. A simple test of significance is available for application to the data in Problem V.6. This is the "sign test" or "binomial series test" for the case of randomized blocks with two columns (Refs. 10 and 5). The statistic used is the number of positive differences among the differences of the several pairs of individuals. The zero differences are usually divided evenly among the positive and negative ones. Thus: $P_0 = +\text{'s} + \frac{1}{2}0\text{'s}$. The mean number of positive values expected according to the binomial series $(\frac{1}{2} + \frac{1}{2})^{25}$ is

$$\begin{aligned} \mu &= np = 12.5 \\ \sigma &= \sqrt{npq} = .5\sqrt{n} = 2.5 \\ X &= \frac{n(P_0 - .50)}{.5\sqrt{n}} \\ &= \frac{25(\frac{18}{25} - .50)}{.5\sqrt{25}} = \frac{25(.72 - .50)}{2.5} = 2.20 \end{aligned} \qquad (5.09)$$

X may be referred to a normal scale (Table I, Appendix), from which it is found that $P = .0278$, or $P < .05$. The hypothesis of no difference between the two groups as revealed by the differences in signs is rejected at the 5 per cent level.

The method differs from the most reliable *t*-test in using only the information in the sign as compared with the total available information in the actual values used by the latter. The former method may be shown to be 62 per cent as efficient as the latter; that is, 62 pairs using the t-test would give as precise results as 100 pairs in using the sign test.[3]

Problem V.8. The significance of the difference between percentages. There is frequent need to determine the significance of the difference between two percentages. Take, for instance, the following problem:

According to one investigator, 67 of an unselected sample of 793 males and 3 of an unselected sample of 232 females from the same United States Caucasoid population were color-blind. Is this evidence of a sex difference in this trait? The hypothesis to be tested is

$$H_0: p_1 = p_2 = p$$

[2] For further discussion of the efficiency of this experimental design see page 292.
[3] For meaning of "efficiency," see page 105.

The maximum likelihood estimate of p is

$$p_0 = \frac{t_1 + t_2}{n_1 + n_2}$$

where t_1 is the number of color-blind individuals in the male sample; t_2, in the female sample.

$$p_0 = \frac{67 + 3}{793 + 232}; \qquad q_0 = 1 - \frac{67 + 3}{793 + 232}$$

$$\left.\begin{aligned} X &= \frac{\dfrac{t_1}{n_1} - \dfrac{t_2}{n_2}}{\sqrt{p_0 q_0 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \\ &= \frac{\frac{67}{793} - \frac{3}{232}}{\sqrt{\left(\frac{70}{1025}\right)\left(\frac{955}{1025}\right)\left(\frac{1}{793} + \frac{1}{232}\right)}} = 3.8 \end{aligned}\right] \tag{5.10}$$

Referred to the normal scale, it is found that $P < .01$. Therefore, the hypothesis is rejected; that is, there is a significant sex difference in color-blindness in the population sampled.[4]

Problem V.9. The significance of the difference between the absolute variabilities of two groups. The following problems illustrate the method of testing the significance of the difference between two variances:

(a) From the measurements of heights in centimeters of 2518 boys and 2538 girls, both groups fourteen years of age, the sum of squares of the deviations for the former was 189,811.641552 and for the latter 114,896.931496. Is there a significant difference in absolute variability?

The calculations are carried out in Table 22.

TABLE 22
THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO ESTIMATES OF VARIANCE

Sex	Degrees of freedom	Sum of squares	Mean square	Log mean square	$\frac{1}{n}$
Male Female	2517 2537	189,811.641552 114,896.931496	75.412 45.306	4.3226 3.8133	.0003973 .0003942

Diff: 0.5093; Sum: .0007915

The mean squares are obtained by dividing the sum of squares by the degrees of freedom. The difference of the logarithms is 0.5093, so z is 0.2546. The variance of z is one-half the sum of the last column, or .0003957; the standard error of z is .0199.

$$\frac{z}{\sigma_z} = 12.8$$

[4] The χ^2 test is an exact test for this problem.

Referred to the normal scale, we find:

$$P < .001$$

Therefore, there is difference in the variability of the two sexes.[5]

(b) Two samples of boys were available in a city school system. One sample of 121 boys of twelve years of age had a mean weight of 72.7 lb. Fifteen years later, another sample of 61 boys twelve years of age from the same school had a mean weight of 77.74 lb. The mean square of the weights $(\text{lb})^2$ of the first sample was 141.60 and that of the second sample, 95.756. Is the difference in variability significant?

The calculations for the test of significance are set forth in Table 23.

TABLE 23
THE z-TEST OF THE SIGNIFICANCE OF THE DIFFERENCE BETWEEN TWO VARIANCE ESTIMATES

Sample	Degrees of freedom	Weight mean square $(\text{lb})^2$	$\frac{1}{2}\log_e$ (mean square)
1	120	141.600	2.4765
2	60	95.756	2.2809

$z_0 = 0.1956$

We enter the z-table of Fisher (Ref. 10) with $n_1 = 120$ and $n_2 = 60$ and by interpolation find that $z_{.05} = .1917$. We could enter Table IV, Appendix, of the variance ratio, F, with $n_1 = 120$, $n_2 = 60$, and $F_0 = \frac{141.60}{95.756} = 1.479$, and find $F_{.05} = 1.472$.

Since z_0 is slightly larger than $z_{.05}$, or F_0 slightly larger than $F_{.05}$, we conclude that the difference in variability in weight between the two groups of boys is significant (at the 5 per cent level).

Problem V.10. To test the homogeneity of a set of estimated variances. The statistical analysis of data often involves the calculation of a number of estimated variances and the testing of whether the sample estimates are significantly different. Three tests of homogeneity of variability are described here.

Neyman and Pearson (Ref. 16) used the criterion L_1, the ratio of a weighted geometric to a weighted arithmetic mean of the mean squares from which the variances were estimated, in order to test the hypothesis, H_1:

$$\sigma_1 = \sigma_2 = \cdots \sigma_k = \sigma$$

This is the test that these k independent samples have been drawn from normal populations having a common standard deviation.

[5] The near normality of z for large and equal values of n_1 and n_2 (see page 55) has been the basis of the test used here. Either the z-test or the variance ratio F could have been used.

Welch (Ref. 24) indicated how L_1 could be generalized and how the weighting for the different sums of squares could be modified. Nayer (Ref. 15) computed tables of the 5 per cent and 1 per cent probability levels for L_1 for the case of equally sized samples. He also considered how far, in the case of unequally sized samples, the probability levels for L_1 might be obtained from his tables. Nair (Ref. 14) investigated the form of the true distribution of L_1.

The test presented here is based on the modified formula of Welch and the use of Nayer's tables of the L_1-distribution.

Welch's equation is

$$L_1 = \prod_s \left(\frac{N}{n_s}\right)^{\frac{n_s}{N}} \prod_s \left\{\frac{\theta_s}{\Sigma\theta_s}\right\}^{\frac{n_s}{N}} \tag{5.11}$$

where $s = 1, 2, \cdots, k$; Π denotes product; Σ denotes summation; n_s, the number of individuals within the sth sample; N, the number of individuals in all the samples; and θ_s is the sum of squares of the errors or the residual of a sample. In the case considered here,

$$\theta_s = \sum_i (X_{si} - \bar{X}_s)^2$$

where X_{si} represents the value of the variate for the ith individual in the sth sample and $\bar{X}_s$ represents the mean of the sth sample.

Nayer's tables are entered with k, the number of samples, and $\bar{n} = \frac{N}{k}$, the average sample size. Hartley (Ref. 12) later indicated that the geometric rather than the arithmetic mean should be used when an average of unequally sized samples is needed. In using L_1-tables, rejection of the hypothesis, H_1, is indicated when the obtained L_1 is equal to or less than the tabled values of L_1 at the respective 1 or 5 per cent level (Table V, Appendix).

The second test of homogeneity of variances was given by Bartlett (Ref. 2). He suggested a test analogous to the L_1 test in which the sums of squares are weighted with the appropriate number of degrees of freedom instead of with the number of observations as in the Neyman-Pearson criterion. Thus where s_t^2 is the unbiased estimate of σ_t^2 based on a sum of squares having v_t degrees of freedom, and there are k independent estimates, the test function is

$$-2 \log_e \mu = N \log_e \left[\sum_{t=1}^{k} \frac{(v_t s_t^2)}{N}\right] - \sum_{t=1}^{k} (v_t \log_e s_t^2) \tag{5.12}$$

$$N = \sum_{t=1}^{k} (v_t)$$

and natural logarithms to the base e are used. Where none of v_t's are too small, $-2 \log_e \mu$ is distributed approximately as χ^2 with $k - 1$ degrees of freedom if the $\sigma_t^2 (t = 1, 2, \cdots, k)$ have a common value. Bartlett gave a corrective factor, C, for small samples:

$$C = 1 + \frac{1}{3(k-1)} \left\{ \sum_t \frac{1}{v_t} - \frac{1}{N} \right\} \tag{5.12a}$$

He indicated that the quantity $\dfrac{-(2 \log_e \mu)}{C}$ followed approximately the same χ^2 distribution.

Bishop and Nair (Ref. 4) demonstrated that even in using the correction factor C, the χ^2 approximation is not altogether satisfactory if some of the degrees of freedom, v_t, are 1, 2, or 3. Later, Hartley (Ref. 12) derived a method of approximating to the distribution of Bartlett's $-2 \log_e \mu$, which was shown to be sufficiently accurate to permit the degrees of freedom to drop to 2 with a fair approximation even if some of the variance estimates based on 1 degree of freedom are among the k-values. In Hartley's method the probability integral is represented as a weighted mean of χ^2 integrals. Thompson and Mennington (Ref. 23) have published tables of the criterion called M, based on Hartley's approximation.

We shall illustrate the three tests of homogeneity of the variances by applying them to the same set of data.

In Table 24, column three is given a set of five estimates of variance, calculated from five samples of intelligence test records of pupils in five different grades of a given school. It is desired to test whether or not there are any real grade differences in the test score dispersion of the pupils. To this end the calculations in obtaining the value of the criterion M_0 are set forth as shown in the table.

TABLE 24

CALCULATIONS FOR OBTAINING THE VALUE OF THE CRITERION FOR BARTLETT'S TEST OF THE HOMOGENEITY OF ESTIMATED VARIANCES

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Grade t	No. of pupils n_t	Intelligence variance s_t^2 (score^2)	v_t	$\log_e s_t^2$	$v_t \log_e s_t^2$	$\frac{1}{v_t}$
3	35	59.5345	34	4.08656	138.94304	0.02941
4	37	98.4369	36	4.58942	165.21912	0.02777
5	35	105.1378	34	4.65527	158.27918	0.02941
6	36	138.3325	35	4.92966	172.53810	0.02857
7	37	39.4520	36	3.67509	132.30324	0.02702
Total	180		175 = N		767.28268	0.14218

We obtain further:

$$\Sigma v_t s_t^2 = 15{,}404.4960$$

$$\frac{\Sigma(v_t s_t^2)}{\Sigma v_t} = \frac{15{,}404.4960}{175} = 88.0257$$

$$\log_e \left\{\frac{\Sigma v_t s_t^2}{\Sigma v_t}\right\} = 4.47763$$

Following (5.12), we obtain

$$-2 \log_e \mu = M_0 = 175 \times 4.47763 - 767.28268 = 16.3026$$

Entering Table VII of the 1 per cent points of the M-distribution (Ref. 23) with $k = 5$ it is found that all entries opposite $k = 5$ are less than 16.3026. Without further calculation, therefore, it may be concluded that $M_0 = 16.3026$ is significant at the 1 per cent level. We may infer that a significant difference exists in the intelligence dispersion, as measured by this test, among the five grades.

Since the tables of the M distribution are not as yet readily accessible, the test may be made by application of (5.12) and (5.12a). Thus:

$$\sum_{t=1}^{k} \frac{(v_t s_t^2)}{N} = \frac{15{,}404.4960}{175} = 88.0257$$

$$N \log_e \frac{\sum_{t=1}^{k} (v_t s_t^2)}{N} = 175(4.47763) = 783.58525$$

$$\sum_{t=1}^{k} (v_t \log_e s_t^2) = 767.28268$$

$$C = 1 + \frac{1}{3(5-1)}\left(\frac{1}{34} + \frac{1}{36} + \frac{1}{34} + \frac{1}{35} + \frac{1}{36} - \frac{1}{175}\right) = 1.01144$$

$$\chi_0^2 = \frac{(783.58525 - 767.28268)}{1.01144} = 16.118$$

We enter the χ^2-table (Table III, Appendix) with $k - 1$, or 4 degrees of freedom. We find that our obtained value χ_0^2 is larger than the table value $\chi^2 = 13.277$ at the one per cent point. Therefore, we reject the hypothesis, H_0, and conclude that a significant difference exists in the variability of intelligence test scores among these five grades.

It may be pointed out here that Bartlett's test would appear advantageous in comparison with the L_1-test when the size of the samples is much larger than $n = 60$ (the limit of finite values as given in the Nayer table) and an interpolation between 60 and infinity needs to be made. Since the range of the L_1-values is only from 0 to 1, the test is not highly sensitive.

The tables of the M distribution may encourage the use of Hartley's approximation, which is likely to be more convenient as well as slightly more accurate.

We now apply the L_1-test, which is made as follows, to the same data as in Table 24.

We calculate first the value log L_1 from Formula (5.11). The calculations are set forth in Table 25.

TABLE 25

THE CALCULATION OF LOG L_1 FOR THE L_1-TEST OF HOMOGENEITY OF VARIANCES

n_s	f_s	$\log n_s$	$n_s \log n_s$	θ'_s	$\log \theta'_s$	$n_s \log \theta'_s$
35	34	1.5441		2,024.1714	3.3062	
37	36	1.5682		3,543.7297	3.5495	
35	34	1.5441		3,574.6857	3.5532	
36	35	1.5563		4,841.6389	3.6850	
37	36	1.5682		1,420.2703	3.1524	
$N = 180$	175	$\sum_s n_s \log n_s = 280.1606$		15,404.4960	$\sum_s n_s \log \theta'_s = 620.7093$	

$$\begin{aligned}
\log L_1 &= \log N - \frac{1}{N}\sum_s n_s \log n_s + \frac{1}{N}\sum_s n_s \log \theta'_s - \log\left(\sum_s \theta'_s\right) \\
&= \log 180 - \tfrac{1}{180}(280.1606) + \tfrac{1}{180}(620.7093) - \log\,(15{,}404.4960) \\
&= 2.25527 - 1.55645 + 3.44838 - 4.18769 \\
&= \bar{1}.95951
\end{aligned}$$

We find that L_1 corresponding to the logarithm $\bar{1}.95951$ is .911.

$$k = 5; \qquad \text{harmonic mean of } f_s = \frac{5}{\left(\frac{1}{34} + \frac{1}{36} + \frac{1}{34} + \frac{1}{35} + \frac{1}{36}\right)} = 34.98$$

We enter Nayer's tables (Table V, Appendix) with $k = 5$ and $f = 35$ and note that our value, .911, is less than the interpolated 1 per cent value of L_1. Therefore, we reject the hypothesis and infer that there is a real difference in variability in the intelligence test scores among the five grades.

Problem V.11. The significance of the difference between two correlation coefficients. The following product-moment coefficients of correlation were obtained between scores on two examinations in algebra administered at the end of the school year in May and at the beginning of the next school year in September. These results were obtained by a mathematics teacher in two different schools:

School A:	$r_1 = .73;$	$n_1 = 59$
School B:	$r_2 = .62;$	$n_2 = 48$

Is there a significant difference between the two correlation coefficients, r_1 and r_2?

$$\begin{aligned} z_1' &= \frac{1}{2}\log_e \frac{1 + r_1}{1 - r_1} \\ &= \frac{1}{2}[\log_e (1 + .73) - \log_e (1 - .73)] \\ &= .9285 \\ z_2' &= \frac{1}{2}\log_e \frac{1 + r_2}{1 - r_2} \\ &= \frac{1}{2}[\log_e (1 + .62) - \log_e (1 - .62)] \\ &= .725 \\ X &= \frac{z_1 - z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}} \\ &= \frac{.9285 - .725}{\sqrt{\frac{1}{56} + \frac{1}{45}}} = .45 \end{aligned}$$

We enter the normal table (Table I, Appendix) and find that for $z = .45$, $P > .05$. We, therefore, conclude that there is no significant difference between the two correlation coefficients.

Problem V.12. The significance of the difference between correlation coefficients determined on the same sample. The following product-moment coefficients of correlation were obtained in a class in college biology, consisting of 73 students.

$r_{y1} = .30$, the correlation coefficient between scores on a test on vocabulary (1) and scores on a test for interpreting various situations dealing with states of health and disease (y);

$r_{y2} = .42$, the correlation coefficient between scores on a test of biological principles (2) and scores on test (y);

$r_{12} = .603$, the correlation coefficient between tests (1) and (2).

The problem is to test the significance of the difference between the correlation coefficients, r_{y1} and r_{y2}. Since these correlations were obtained on the same sample the procedure described on page 54 is followed:

$$\begin{aligned} F &= \frac{(r_{y1} - r_{y2})^2(N - 3)(1 + r_{12})}{2(1 - r_{12}^2 - r_{y1}^2 - r_{y2}^2 + 2r_{12}r_{y1}r_{y2})} \\ F_0 &= \frac{(.30 - .42)^2(73 - 3)(1 + .603)}{2[1 - (.603)^2 - (.30)^2 - (.42)^2 + 2(.603)(.30)(.42)]} \\ &= 1.55 \end{aligned}$$

We enter the table of F (Table IV, Appendix) with $n_1 = 1$ and $n_2 = 70$. We find that our value, 1.55, is less than $F_{.05} = 3.98$; hence F_0 is not significant. We conclude that there is no significant difference in the two correlation coefficients.

Problem V.13. To test the significance of a regression coefficient. In a simple regression of one independent variable, an important test is whether the regression coefficient is significantly different from zero, or the test of the hypothesis that there is no regression of y on x in the population sampled. For the required test the theoretical model against which the experimental results may be compared is the t-distribution. The value of t_0 is calculated from the sample and the table of t is entered with $n = N - 2$ to determine the probability of getting a value of t greater than or equal to $\pm t_0$ in repeated sampling. Here t_0 is the ratio of the regression coefficient, b_{yx}, to the standard error of b_{yx}; that is,

$$t_0 = \frac{b_{yx}}{\sigma_b} \tag{5.13}$$

The standard error of b is given by

$$\sigma_b = \frac{\sigma_{y.x}}{\sqrt{\Sigma x^2}} \tag{5.14}$$

where $\sigma_{y.x}$, the standard error of estimate, is obtained from

$$\sigma_{y.x} = \sqrt{\frac{\Sigma(Y_0 - Y_E)^2}{N-2}} = \sqrt{\frac{\Sigma y^2 - \frac{(\Sigma xy)^2}{\Sigma x^2}}{N-2}} \tag{5.15}$$

in which Y_0 is the observed value of Y and Y_E is the value of Y estimated from the regression equation. The number of degrees of freedom is $N - 2$, since two statistics are estimates of two different parameter values in the regression equation: $Y_E = a + bx$. The calculations and procedures are illustrated in determining the regression coefficient and in testing its significance in Table 26.

This problem was that of setting up a regression equation for the purpose of predicting a knowledge of one character Y, from a knowledge of a second character X. In this case it was desired to predict the score of an individual on one form of an examination from his score on the second form. The prediction equation is

$$Y_E = \bar{Y} + \frac{\Sigma xy}{\Sigma x^2}(X - \bar{X}) \tag{5.16}$$

This equation is called the *regression equation* for estimating Y from X. It is fitted to the observational data by the method of least squares.

In this regression problem, it is necessary to run two tests of significance: (1) for the regression coefficient b_{yx} and (2) for the mean of the dependent variable, $\bar{Y}$.

The test of significance for the regression coefficient, b_{yx}, is given by $t = \frac{b_{yx}}{s_b}$.

In our problem the values are

$$t_0 = \frac{.9873}{.0729} = 13.5$$

We enter the table of t with $n = N - 2 = 25 - 2 = 23$ and find that $P < .001$. Therefore, the regression coefficient is highly significant.

TABLE 26

CALCULATIONS FOR SETTING UP THE REGRESSION EQUATION BETWEEN THE SCORES ON TWO FORMS OF A TEST

Individual	X	Y	X'	Y'	X'^2	Y'^2	$X'Y'$
1	46	52	− 4	2	16	4	− 8
2	38	38	−12	−12	144	144	144
3	64	63	14	13	196	169	182
4	73	65	23	15	529	225	345
5	61	58	11	8	121	64	88
6	34	33	−16	−17	256	289	272
7	57	49	7	− 1	49	1	− 7
8	66	63	16	13	256	169	208
9	25	24	−25	−26	625	676	650
10	30	26	−20	−24	400	576	480
11	45	33	− 5	−17	25	289	85
12	73	71	23	21	529	441	483
13	45	48	− 5	− 2	25	4	10
14	55	63	5	13	25	169	65
15	66	70	16	20	256	400	320
16	49	46	− 1	− 4	1	16	4
17	64	65	14	15	196	225	210
18	45	46	− 5	− 4	25	16	20
19	61	62	11	12	121	144	132
20	52	46	2	− 4	4	16	− 8
21	67	68	17	18	289	324	306
22	59	53	9	3	81	9	27
23	55	55	5	5	25	25	25
24	51	52	1	2	1	4	2
25	50	48	0	− 2	0	4	0
Total	1331	1297	81	47	4195	4403	4035

X and Y are scores on two tests.
$X' = X - 50$ and $Y' = Y - 50$.

$$\bar{X} = 50 + \tfrac{81}{25} = 53.24$$
$$= \text{mean of } X \text{ scores}$$
$$\bar{Y} = 50 + \tfrac{47}{25} = 51.88$$
$$= \text{mean of } Y \text{ scores}$$

Let $x = X - \bar{X}$ and $y = Y - \bar{Y}$

$$\Sigma x^2 = \Sigma(X - \bar{X})^2 = \Sigma X'^2 - \frac{(\Sigma X')^2}{25}$$
$$= 4195 - \frac{(81)^2}{25}$$
$$= 4195 - 262.44 = 3932.56$$
$$\Sigma y^2 = 4403 - \frac{(47)^2}{25}$$
$$= 4403 - 88.36 = 4314.64$$
$$\Sigma xy = \Sigma X'Y' - \frac{(\Sigma X')(\Sigma Y')}{25}$$
$$= 4035 - \frac{(81)(47)}{25}$$
$$= 4035 - 152.28 = 3882.72$$

Regression Equation

$$Y_E = \bar{Y} + \frac{\Sigma xy}{\Sigma x^2}(X - \bar{X})$$
$$= 51.88 + \frac{3882.72}{3932.56}(X - 53.24)$$
$$= 51.88 + .9873(X - 53.24)$$
$$= 51.88 + .9873X - 52.56$$
$$= .9873X - 0.68$$

Significance of Regression Coefficient

$$\sigma_b = \frac{\sigma_{y \cdot x}}{\sqrt{\Sigma x^2}} = \frac{4.57}{\sqrt{3932.56}} = \frac{4.57}{62.7}$$
$$= .0729 = \text{standard error of the regression coefficient}$$
$$t = \frac{.9873}{.0729} = 13.5 \sim P < .01$$

TABLE 26 (*Continued*)

Standard Error of Estimate

$$\sigma_{y\cdot x} = \sqrt{\frac{\Sigma(Y_0 - Y_E)^2}{N - 2}} = \sqrt{\frac{\Sigma y^2 - \dfrac{(\Sigma xy)^2}{\Sigma x^2}}{N - 2}}$$

$$= \sqrt{\frac{4314.64 - \dfrac{(3882.72)^2}{3932.56}}{23}}$$

$$= \sqrt{\frac{4314.64 - 3833.51}{23}} = \sqrt{\frac{481.13}{23}} = \sqrt{20.9189}$$

$= 4.57 =$ standard error of estimate

Test of Significance of $\bar{y}$

$$t = \frac{(\bar{y} - \alpha)\sqrt{n}}{s}$$

where $s = \sqrt{\dfrac{\Sigma(Y_0 - Y_E)^2}{n - 2}}$

$n = n' - 2$

A simpler alternative test of the significance of the regression coefficient can be made, where the correlation coefficient, r_{xy}, is available and under the conditions indicated below.

When the regression of y on x is linear and the arrays of y are normal and homoscedastic[6] (see Ref. 11), the t-test affords the exact test of the significance of the deviation of a sample regression coefficient from any hypothetical value (specified by the hypothesis tested) divided by an estimate of its standard error, considered as a random sample of similar estimates in repeated samples with the same values of x.

When the hypothesis under test is that the population value, ρ, is zero and when the distribution of X is continuous, the t-test for b_{yx} is also an exact test for the sample correlation coefficient, r:

$$t = \frac{b}{s_b} = \frac{r\sqrt{N - 2}}{\sqrt{1 - r^2}} \tag{5.17}$$

This equality is illustrated by calculating r for the set of data in Table 26. Thus, with $r = .94$,

$$t_0 = \frac{.9873}{.0729} = \frac{.94\sqrt{23}}{\sqrt{1 - (.94)^2}} = 13.5$$

Entering the t-table with $n = 23$, it is observed that $P < .001$. Therefore, the observed value of b (or r) is highly significant.

For a test of the hypothesis that two regression coefficients b_1 and b_2, obtained from two random samples of sizes N_1 and N_2, are from the same population, t_0 is given by

$$t_0 = \frac{b_1 - b_2}{s'\left(\dfrac{1}{\Sigma x_1^2} + \dfrac{1}{\Sigma x_2^2}\right)^{\frac{1}{2}}} \tag{5.18}$$

where $s'^2 = \dfrac{\Sigma(Y_1 - Y_{E_1})^2 + \Sigma(Y_2 - Y_{E_2})^2}{N_1 + N_2 - 4}$

$$(n = N_1 + N_2 - 4) \tag{5.19}$$

[6] For tests of linearity and homoscedasticity, see page 241.

As an alternative, the significance of the difference between the two correlation coefficients could be tested as in Problem V.11.

Problem V.14. The significance of the mean of the dependent variable in a simple regression equation. The same set of data used in the preceding problem may be presented to illustrate the test of significance of the second estimate in the regression equation, the estimate of any hypothetical value α. In this case the t-distribution may also be used. The sample value of t_0 is

$$t = \frac{(\bar{Y} - \alpha)\sqrt{N}}{s} \qquad (5.20)$$

where

$$s = \sqrt{\frac{\Sigma(Y_0 - Y_E)^2}{N - 2}} \qquad (5.21)$$

Then for this sample and where α is specified as zero, t_0 becomes

$$t_0 = \frac{51.88\sqrt{25}}{4.57} = 56.7 \qquad (n = N - 2)$$

Obviously, $P < .001$.

Applications of the Chi-Square Model. The chi-square (Ref. 21) model has wide application in statistics, particularly as a test of significance in dealing with enumerative data so characteristic of the study of attributes. It is appropriate for testing whether a set of observed values differs significantly from those which would occur if some specified hypothesis were true. One general method of testing such a hypothesis is to work out results which would be expected theoretically and then to compare these with the observations.

Problem V.15. To test the effectiveness of principles of classification. We may have individuals classified by two characteristics and wish to test the hypothesis that the characteristics are independent or that the principles of classification are independent.

In applying the χ^2-test to two or more classifications, usually the statistical hypothesis under test is that the two characteristics upon which the individuals have been classified are independent of one another, and then the truth or falsity of the hypothesis is tested. This procedure is equivalent to determining whether a set of obtained values differs significantly from those which would result if only chance factors were in operation.

In the following example the χ^2-test is applied to a 2 × 2-fold contingency table (Table 27).

Here 366 twins have been classified on the basis of two characteristics according to (1) their genetic constitution, that is, according to whether they are identical or fraternal twins, and (2) the presence or absence of mental deficiency. The numbers of identical and fraternal twins are recorded in the marginal totals in the last column of the table of observed values. The number of concordant and disconcordant twins with respect to mental deficiency is given in the marginal totals in the last row of the table.

The χ^2-test is applied to determine the independence of these two factors. The geneticist or psychologist might state the problem thus: Assuming that the data are accurate, homogeneous, and unselected, with what frequency could so large a disproportion between the two classes of twins arise if the same causes leading to mental deficiency had been operative on the two?

TABLE 27

CONCORDANCE AND DISCONCORDANCE IN IDENTICAL AND FRATERNAL TWINS FOR MENTAL DEFICIENCY (After Rosanoff, Ref. 20)

Type	Observed values		
	Number concordant	Number disconcordant	Total
Identical twins	115(a)	11(b)	126
Fraternal twins	128(c)	112(d)	240
Total	243	123	366

Type	Expected values		
	Number concordant	Number disconcordant	Total
Identical twins	83.66(a)	42.34(b)	126
Fraternal twins	159.34(c)	80.66(d)	240
Total	243.00	123.00	366

The number of observations to be expected in each cell where only chance factors are operative can be calculated from the total frequency in this way: Multiply the total number of identical twins, 126, by the total number of concordant twins, 243, that is, 126 × 243 = 30,618, and divide this product by the total number of twins in the sample, 366, that is, 30,618/366 = 83.66. The expected number in the other cells of the tables can be calculated in the same way. This need not be done, however, in a 2 × 2 table. Since the marginal totals are fixed, the expected values for only one cell need be calculated, the others being filled in by subtraction. Thus, the expected value for cell *b* is 126 − 83.66 = 42.34; that for cell *c*, is 243 − 83.66 = 159.34; and that for cell *d* is 123 − 42.34 = 80.66.

χ^2 is given by the formula

$$\chi^2 = \sum \frac{(f_0 - f_t)^2}{f_t} \tag{5.22}$$

where f_0 stands for observed frequency and f_t for expected frequency The square of the differences between the observed and expected values is divided by the expected value for each cell. These quotients are summed to give χ^2.

The calculations for the above data are presented in Table 28.

TABLE 28

THE CALCULATION OF χ^2 FOR THE DATA IN TABLE 27

Cell	f_0	f_t	$(f_0 - f_t)$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
a	115	83.66	31.34	982.1956	11.74
b	11	42.34	−31.34	982.1956	23.20
c	128	159.34	−31.34	982.1956	6.17
d	112	80.66	31.34	982.1956	12.18
Total	366	366.00	00.00	$\chi_0^2 = 53.29$	

The calculated value χ_0^2 is used to determine the probability of getting, on a random sample, the value of χ^2 equal to or higher than χ_0^2 in repeated sampling. The alternative is the probability that the difference between the observed and expected values may be attributable to chance alone. This probability is obtainable from Table III, Appendix, Distribution of χ^2. The number of degrees of freedom with which the table is entered is in this problem equal to 1, since it was observed that only one of the cell frequencies could be filled in independently. When this quantity is specified, the other cells can be filled in by using the marginal totals.

Therefore, we enter the χ^2-table with a value of $\chi_0^2 = 53.29$ and $n = 1$. It is noted that for values of χ^2 greater than 10.827 the probability that the differences between the observed and obtained frequencies could have arisen by chance is $< .001$. The table does not give the value of P for a value of $\chi^2 = 53.29$. The probability, however, is much less than 1 in 1000. Hence it may be concluded that the system of classification used in this problem was effective, or that the two basic characteristics, type of twin and mental deficiency, are associated.

It may be pointed out that in a 2 × 2 table the value of χ^2 could have been obtained directly from the formula

$$\chi^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(a + c)(b + d)} \tag{5.23}$$

In our problem,

$$\chi_0^2 = \frac{[(115)(112) - (11)(128)]^2(366)}{(126)(240)(243)(123)} = 53.29$$

The correction for continuity, devised by Yates (Ref. 25), is useful for extending the application of χ^2-test of significance to contingency tables with small frequency, that is, to data in which the expectations are small. Cochran (Ref. 6) has presented and illustrated the principles involved in correcting for continuity on some applications of χ^2.

The process of calculating χ^2 for a 2 × 2 table can be extended to the general case of the $r \times c$ contingency table. In general, in a table of r rows and c columns the number of degrees of freedom in χ^2 is $(r - 1)(c - 1)$. Bartlett (Ref. 1) devised a method of calculating χ^2 for multiple-dichotomous tables, that is, those of the form 2^N. Norton (Ref. 17) presented and illustrated a method of successive approximation for obtaining the R departures from expectation in a complex contingency table of the form $2^N \times R$.

Problem V.16. To test the homogeneity of two or more frequency distributions. A useful application of the χ^2-test is in testing the hypothesis that two or more frequency distributions could have come from the same homogeneous population. This is a more stringent test than those tests of the significance between certain summary statistics of the distributions, since by it the distributions are compared in all respects. Furthermore, it is possible to separate the contributions to χ^2 of the individual degrees of freedom, and so to test the distributions by parts.

The following example illustrates the case where there are two distributions and n' classes with $n' - 1$ degrees of freedom. The method of calculating χ^2 devised by Brandt and Snedecor (Ref. 11) is followed.

The two samples are distributions of two groups of freshmen entering a particular college of the University of Minnesota classified according to college aptitude-test rating. One distribution, of 475 students, presented two units of high-school mathematics; the other, of 111 students presented three units of high-school mathematics at the time of entrance. We wish to test the hypothesis that these two samples are from the same homogeneous population with respect to aptitude as measured, or whether there is a significant difference between the two distributions.

If we denote the column of frequencies of the group with two units by a', that of the group with three units by a, the value of χ^2 is given by the formula

$$\chi^2 = \frac{1}{\bar{p}\bar{q}}\left(\sum (ap) - n\bar{p}\right) \tag{5.24}$$

where $p = \dfrac{a}{(a + a')}$

$$\bar{p} = \frac{n_1}{n_1 + n_2}$$

The calculations of χ^2 for the test of significance of the homogeneity of the two frequency distributions are given in Table 29.

For a $\chi_0^2 = 30.96$ with $n = 9$, we enter the χ^2-table and find that for

TABLE 29

CALCULATION OF χ^2 FOR TWO FREQUENCY DISTRIBUTIONS—ONE WITH TWO UNITS OF HIGH-SCHOOL MATHEMATICS, THE OTHER WITH THREE, GROUPED ACCORDING TO PERCENTILE RANKS ON THE COLLEGE APTITUDE TEST

Class intervals in percentile ranks	Units of high-school mathematics		$P = \frac{a}{a + a'}$	aP
	Two (a')	Three (a)		
91–100	18	10	.357143	3.571430
81– 90	33	12	.266666	3.199992
71– 80	39	14	.264151	3.698114
61– 70	43	3	.652087	1.956261
51– 60	39	13	.250000	3.250000
41– 50	51	3	.055550	0.166650
31– 40	47	12	.203390	2.440680
21– 30	66	14	.175000	2.450000
11– 20	68	8	.105263	0.842104
0– 10	71	22	.236559	4.204298
Total	475(n_2)	111(n_1)	.189420	25.779529

$\bar{P}$ Σap

$$\chi_0^2 = \frac{1}{(.18942)(.81058)}[25.779529 - (111)(.18942)]$$
$$= 30.96 \sim P < .001; \text{for } n = 9, \chi^2_{.001} = 27.877$$

values of χ^2 greater than 27.877 the divergencies between the observed frequencies in the two distributions could have arisen by chance in less than .001. We do not know the value of P, corresponding to a value of $\chi^2 = 30.96$, but the probability of such a divergence arising by chance is less than 1 in 1000. We may conclude, therefore, that there is a statistically significant difference between the two distributions. The pedagogical conclusion is that groups presenting three units of high-school mathematics are superior on the whole on the College Ability Test to the groups presenting two units.

It is possible to separate the contributions to χ^2 from each of the individual degrees of freedom, and so to test the distributions by parts.

For 4 degrees of freedom the calculations for χ^2 are

Percentile ranks on College Aptitude Test	Two units	Three units	Total	P
81–100	51	22	73	.301370
61– 80	82	17	99	.171717
41– 60	90	16	106	.150943
21– 40	113	26	139	.187050
1– 20	139	30	169	.177515
Total	475	111	586	.189420($\bar{P}$)

$$\chi_0^2 = 7.3438$$
$$\sim P > .05$$

For 1 degree of freedom:

P.R.C.A.T.	Two units	Three units	Total	P
Above 80 P.R.	51	22	73	.301370
80 and below	424	89	513	.173490
Total	475	111	586	.189420($\bar{P}$)

$$\chi_0^2 = 6.8066$$
$$\chi_{.01}^2 = 6.635$$
$$P\chi_0^2 < .01$$

The portion of the distribution contributing the most to the differences is, accordingly, in the highest percentile ranks, or 81–100.[7]

Problem V.17. To test the agreement between a theoretical and an observed distribution. One general method of testing a statistical hypothesis is to work out the results which would be expected theoretically under the assumption that the hypothesis is true, and then to compare these with the observations. The chi-square test provides an efficient test of the goodness of fit. As an illustration we shall test the hypothesis that a set of data presented by Roberts *et al.* (Ref. 19) is described by a Poisson series.

The data given in Table 30 were obtained in administering the Binet Test (a shortened form) to a group of children who passed all but one

TABLE 30

ADDITIONAL TESTS FAILED ON DOWNARD EXTENSION OF THE BINET SCALE TO A SAMPLE OF 131 CHILDREN

(After Roberts, Ref. 19)

Number of tests failed	Observed frequency, f_0	Expected frequency,* f_t	χ^2 †
0	88	87.41	0.004
1	34	35.37	0.053
2	8	$7.1\dot{6}$	0.070
3	1	$0.9\dot{7}$	
4	0	$0.1\dot{0}$	
5	0	$0.0\dot{1}$	
Total	131	131.02	0.127

* The theoretical distribution is obtained as follows:

[7] The reduction in the χ^2-values with coarser grouping of the data is noted. This result is to be expected with the reduction in the number of degrees of freedom and the corresponding approach to the zero tail of the χ^2-distribution.

TABLE 30 (*Continued*)

1. Calculate the mean number of tests failed: $\bar{X} = \frac{53}{131} = 0.4046$.
2. Calculate the expected frequency. This is done by means of logarithms. Thus:

Quantity	Logarithm	Expected frequency
$n = 131$	2.11727	
$e^m = e^{0.4046}$ $(m \log e) =$		
$(.4046)(.43429) =$	0.17571	
n/e^m	1.94156	87.41
$m = 0.4046$	9.60703 − 10	
mn/e^m	1.54859	35.37
m	9.60703 − 10	
	1.15562	
	0.30103	
$m^{2^2}n/2e^m$	0.85459	7.155
m	9.60703 − 10	
	0.46162	
	0.47712	
$m^{3^3}n/(2)(3)e^m$	9.98450 − 10	0.965
m	9.60703 − 10	
	9.59153 − 10	
	0.60206	
$m^{4^4}n/(2)(3)(4)e^m$	8.98947 − 10	0.0976
m	9.60703 − 10	
	8.59650 − 10	
	0.69897	
$m^{5^5}n/(2)(3)(4)(5)e^m$	7.89753 − 10	0.007898

† Chi-square is determined in the usual manner by calculating $\frac{\Sigma(f_o - f_t)^2}{f_t}$. The classes from 2 through 5 have been grouped because of small frequencies. This grouping could have been done without calculating the theoretical values for the classes beyond the third. The calculations were made to illustrate the method. $\chi_0^2 = 0.127$ with $n = 1$. The corresponding probability value is $.70 < P < 80$. There is 1 degree of freedom, since the sample mean has been used as the parameter of the Poisson distribution and the sample number has been used to calculate the theoretical frequencies.

of a complete year of tests, then by extending the testing downward to determine how many of these pupils failed in one, two, or more tests.

From the calculations it is noted that for a $\chi_0^2 = 0.127$ and with $n = 1$, the corresponding probability is between .70 and .80. Therefore, we may conclude that the Poisson distribution provides a good fit to this set of data.

PROBLEMS

1. The following are two distributions A and B from the Miller Analogies Test. Determine whether they are random samples from the same population.

A					
85	53	43	77	53	79
51	56	99	57	75	67
80	88	76	84	48	86
77	69	84	83	89	77
56	72	71	27	48	90

B					
48	66	50	65	75	48
47	53	99	79	53	79
67	75	76	69	75	62
75	54	69	78	48	96
76	75	76	67	89	84
					92

2. The following data are from an experiment comparing the relative efficacy of two different methods of teaching beginning high-school algebra. There are XXIV pairs of students, paired on the basis of chronological age and a pretest in arithmetic. There were two criteria of achievement: (1) scores on an inventory test, (2) scores on an achievement test. The values under "Exp." and "Con." refer to the experimental and control groups, respectively. Test the null hypothesis in this experiment.

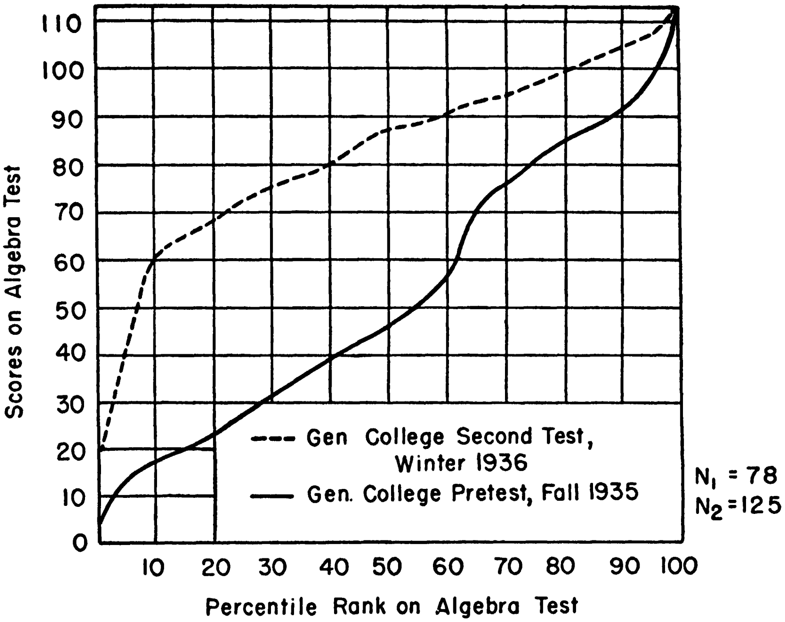
Data for Problem 2

Pairs	Chron. Age		Arith.		Inventory		Achievement	
	Exp.	Con.	Exp.	Con.	Exp.	Con.	Exp.	Con.
I	152	157	99	96	57	57	50	33
II	172	157	87	87	61	67	32	20
III	173	177	85.5	86.5	60	62	24	28
IV	169	166	85	86.5	55	55	28	19
V	160	156	85	86.5	50	50	33	23
VI	168	162	82.5	82	50	50	31	28
VII	171	169	96.5	97.5	56	57	36	28
VIII	160	156	92	91	56	56	43	31
IX	177	171	83	86	60	60	33	21
X	165	161	85.5	86.5	57	56	33	21
XI	164	165	87.5	84.5	57	56	38	27
XII	167	161	84.5	83	56	56	28	28
XIII	171	171	96.5	96	57	57	35	20
XIV	168	169	99.5	99	50	51	42	24
XV	175	177	83	80.5	56	56	29	27
XVI	172	175	93	90	56	56	41	18
XVII	169	170	81.5	79	56	57	35	20
XVIII	161	167	90	87.5	56	56	36	26
XIX	165	171	87.5	87.5	56	56	28	28
XX	174	168	83	85.5	56	56	20	27
XXI	176	175	93	94	51	50	29	29
XXII	170	165	77	79.5	42	50	42	29
XXIII	174	172	77	79.5	56	56	29	24
XXIV	174	170	86	86	50	50	38	30

3. (a) In Problem 2 determine the statistical significance of the differences in achievement of the two groups by only considering the signs of the respective differences between the scores of individual pair members.

(b) Compare the efficiency of the test used in (a) with that of the test used in Problem 2.

4. Determine the significance of the difference of the percentage of those taking the second test, reaching or exceeding the median score of the group taking the pretest in the fall of 1935.



5. For the following distribution calculate (a) The variance from the grand mean; (b) the variance from the sample means.

(c) Note the extent of agreement.

(d) Why is it necessary to pay proper regard to the number of degrees of freedom?

Sample	I	II	III	IV	V	VI	VII	VIII	IX	X
	12	25	18	8	20	21	15	24	28	29
	29	25	22	21	22	19	23	10	25	21
	22	23	17	17	24	12	18	23	14	18
	20	14	23	20	14	11	23	22	20	16
	22	24	11	14	22	14	20	20	29	22

6. A check-up on the reading habits of seventh-grade pupils reveals that 55 per cent of the 558 voluntary readings of one random sample of pupils was mystery and detective, where only 45 per cent of the 122 voluntary readings of another random sample of pupils was of this classification. Is there statistical evidence here that interest in the mystery and detective type of reading is higher in one sample than in the other?

7. In an attitude test administered to an experimental group of 796 students and a control group of 861, Item 306 was answered correctly by 51 in the experimental group and by 47 in the control group. (a) Is there a statistically significant difference between the proportion of the experimental group that answered this item correctly and the proportion of the control group that answered it correctly? What is the statistical hypothesis tested? (b) In Item 35, 37 of the experimental group and 37 of the control group answered this item correctly. Answer the above questions in regard to this item.

8. The following measures were obtained from an examination in personal hygiene for a winter quarter class and for a spring quarter class. Determine the significance of the difference between means. May the variances be assumed equal? What hypothesis is under test? What is the most appropriate test of the hypothesis?

Winter quarter class:
Mean = 20.56
Sum of squares of deviations from the mean = 28,255
Number = 675
Spring quarter class:
Mean = 22.07
Sum of squares of deviations from the mean = 12,535
Number = 350

9. In a given situation $n = 81$, mean $= 40$, and standard deviation $= 8$. If we assume that the standard deviation of the increased number of cases will remain approximately the same as given, what size of sample is necessary to reduce the standard error of the mean to .5?

10. The following data indicate the frequency of intrapair differences in handedness in identical twins and in the handedness of their immediate relatives:

	Identical twins	
	R-R	R-L
Without left-handed relatives	105	25
With left-handed relatives	26	22
Total	131	47

Is the principle of classification effective?

11. Following are two distributions of entering freshmen, the one having had no high-school work in foreign languages, the other having had two or more units in foreign languages. Test the independence of the two distributions as wholes and by parts.

Frequency subgroups, percentile ranks on College Aptitude Test	Units in high-school foreign language	
	None	Two or more
91–100	8	20
81– 90	7	27
71– 80	11	33
61– 70	8	29
51– 60	15	25
41– 50	10	34
31– 40	11	33
21– 30	35	38
11– 20	24	50
1– 10	45	50
	$N_1 = 174$	$N_2 = 339$

12. The following data were obtained from four random samples of entering freshmen on a chemistry aptitude test:

Entering Group	N	$\bar{X}$	s
1938	35	18.66	3.58
1939	48	17.23	4.75
1940	42	18.67	4.95
1941	30	19.53	3.09
Total	155	18.39	4.33

Test the homogeneity of the standard deviations.

13. The following coefficients of correlation were reported between intelligence quotients (X) and chronological ages (Y) for two random samples of students in a course in elementary-school science. Test the significance of the difference between the two correlation coefficients:

Sample 1: $N_1 = 96$ $\quad r_{xy} = -.507$
Sample 2: $N_2 = 66$ $\quad r_{xy} = -.455$

14. The following correlation coefficients were obtained upon a random sample of 74 pupils in the sixth grade of an elementary school:

$$r_{xy} = .55; \qquad r_{xz} = .81; \qquad r_{yz} = .44$$

where x = score on an initial achievement test
y = mental-age score
z = score on a final achievement test.

Test the significance of the difference between r_{xz} and r_{yz}.

15. Test the significance of the differences among the following correlation coefficients reported for the illustrative problem in multiple correlation (see page 332).

$$r_{1y} = .1784 \qquad r_{3y} = .5164 \qquad r_{5y} = .6704$$
$$r_{2y} = .6505 \qquad r_{4y} = .0993$$

References

1. Bartlett, M. S., "Contingency Table Interactions," *Supplement to Journal of the Royal Statistical Society,* Vol. II (1935), pp. 248–252.
2. ———, "Properties of Sufficiency and Statistical Tests," *Proceedings of the Royal Society (London),* Series A, Vol. CLX (1937), pp. 1–273.
3. Behrens, W. U., "Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen," *Landwirtschaftliche Jahrbücher,* Vol. 68 (1929), pp. 807–837.
4. Bishop, D. J., and Nair, U. S., "A Note on Certain Methods of Testing for the Homogeneity of a Set of Estimated Variances," *Supplement* 6 *to Journal of the Royal Statistical Society,* Vol. VI (1939), pp. 89–99.
5. Brandt, A. E., "A Test for Significance in a Unique Sample," *Journal of the American Statistical Association,* Vol. 38 (1933), pp. 434–437.
6. Cochran, W. G., "The χ^2 Correction for Continuity," *Iowa State College Journal of Science,* Vol. XVI (1942), pp. 421–436.
7. Fisher, R. A., "The Asymptotic Approach to Behrens' Integral with Further Tables for the *d*-Test of Significance," *Annals of Eugenics,* Vol. XI (1941).
8. ———, "The Comparison of Samples with Possibly Unequal Variances," *Annals of Eugenics,* Vol. IX (1939), p. 174.
9. ———, "The Fiducial Argument in Statistical Inference," *Annals of Eugenics,* Vol. 6 (1935), pp. 391–398.
10. ———, *Statistical Methods for Research Workers,* 1st ed. London: Oliver and Boyd, 1925, Section 24, Ex. 9.
11. *Idem.,* 4th ed., 1932, p. 90.
12. Hartley, H. O., "Testing the Homogeneity of a Set of Variances," *Biometrika,* Vol. XXXI (1940), pp. 249–255.
13. Hsu, P. L., "Contribution to the Theory of 'Student's' *t*-Test as Applied to the Problem of Two Samples," *Statistical Research Memoirs,* Vol. II (1938), pp. 1–24.
14. Nair, U. S., "The Form of the L_1-Distribution," *Biometrika,* Vol. XXX (1940), pp. 249–255.
15. Nayer, P. P. N., "An Investigation into the Application of Neyman and Pearson's L_1 Test, with Tables of Percentage Limits," *Statistical Research Memoirs,* Vol. I (1936), pp. 38–51.
16. Neyman, J., and Pearson, E. S., "On the Problem of *k*-Samples," *Bulletin de l'Acadḿie Polonaise des Sciences et des Lettres,* Série A (1931), pp. 460–481.
17. Norton, H. W., "Calculation of Chi-Square for Complex Contingency Tables," *Journal of the American Statistical Association,* Vol. 40 (1945), pp. 251–258.
18. Peterson, A. S., "The Evaluation of a One-year Course, the Fusion of Physics and Chemistry with Other Physical Science Courses." Unpublished Ph.D. thesis, University of Minnesota, 1944.
19. Roberts, J. A. F., *et al.,* "Studies on a Child Population," *Annals of Eugenics,* Vol. 8 (1938), pp. 319–36.
20. Rosanoff, A. J., Handy, L. M., and Plesset, I., "The Etiology of Mental Deficiency with Special Reference to its Occurrence in Twins," *Psychological Monograph No. 216* (1937) Princeton N. J.; Published for American Psychological Association.
21. Snedecor, George W., *Statistical Methods,* 4th ed. Ames, Iowa: Collegiate Press, 1946, pp. 83–84.
22. Sukhatme, P. V., "On the Fisher-Behrens Test of Significance for the Difference in Means of Two Normal Samples," *Sankhya,* Vol. 4, Part 1 (1938), p. 39.

23. Thompson, Catherine M., and Merrington, Maxine, "Tables for Testing the Homogeneity of a Set of Estimated Variances," *Biometrika*, Vol. XXXIII, Part IV (1946), pp. 296–304.

24. Welch, B. L., "Some Problems in the Analysis of Regression Among k-Samples of Two Variables," *Biometrika*, Vol. XXVII (1935), p. 145.

25. Yates, F., "Contingency Tables Involving Small Numbers and the χ^2 Test," *Supplement to Journal of the Royal Statistical Society*, Vol. I (1934), pp. 217–235.

CHAPTER VI

THE ESTIMATION OF POPULATION PARAMETERS

The Problem of Estimation. The estimation of characteristics of a population, that is, the estimation of the parameter values of the population, is a fundamental statistical problem. In such a problem we usually begin with an assumption about or knowledge of the mathematical form of the population of which we presume to have a random sample. We do not have a knowledge of the values of one or more parameters in the mathematical form. These values are required for the complete specification of the population.

In general, there are a number of ways of estimating a parameter from sample data, some of which may be better than others. The theory of estimation provides a basis for investigating the conditions which an estimate should fulfill, for determining the best estimate to use under given circumstances, and for comparing the relative effectiveness of different estimates that might be used.

In its most practical form, the problem of estimation is met with by the research worker in his attempt to reduce his original data to a few summary quantities which shall contain all the relevant information, that is, all information which is of use in estimating the values of the parameters. The problem of estimation is closely related to that of distribution, since both arise in the process of reducing data. From the logical standpoint, problems of distribution precede problems of estimation, since knowledge of the random distributions of various alternative statistics, derived from samples of a given size, is basic in the selection of the particular statistic most useful to calculate.

The problem of specification, or the specification of the mathematical form of the distribution of the hypothetical population from which a sample is assumed to have been drawn, completes the theoretical basis upon which depends the solution of the problems which arise in the reduction of data. Although the three problems may be studied separately, evidently they are closely related in the development of statistical methods. Our purpose here is to study especially the problem of estimation. This is the problem of determining how observational data can be best combined to yield the most accurate estimates obtainable of the unknown parameters. Two procedures of estimation are considered: (1) estimation by a point and (2) estimation by an interval.

In order to judge whether one particular estimate or a group of estimates is better than others, criteria are needed. Three criteria have been advanced: (1) consistency, (2) efficiency, and (3) sufficiency. Statistics which satisfy these criteria are known as *optimum estimates* or *optimum statistics* (Ref. 13).

CHARACTERISTICS OF GOOD ESTIMATES

In order to be *consistent,* the value of a statistic must approach more and more closely the estimated parameter as the sample size is indefinitely increased. Such a value is a function of the observations, which converges stochastically to a population parameter as the sample number approaches infinity. An *efficient* estimate is one whose sampling distribution tends to the normal law with the least possible standard error as the number of observations is increased. Efficiency requires that the variance of the estimate (at least for large samples) should not exceed that of any other consistent statistic estimating the same parameter. The square of the ratio of the minimum standard error to the standard error of another estimate (also normally distributed in the limit) gives a measure of the relative efficiency of the second estimate. The criterion of *sufficiency* is satisfied by a statistic when no other statistic calculable from the same sample can supply any additional information regarding the parameter under estimation. A sufficient statistic is inevitably also 100 per cent efficient, since it incorporates the whole of the information available in the sample in regard to a given parameter.

The Measurement of Amount of Information. It is apparent that these criteria for judging the goodness of estimates require the knowledge of the amount of information that is available in any sample relevant to the population parameter under estimation. Fisher (1921, 1925) showed how to measure the quantity of information provided by the observational data, relevant to the value of any particular unknown quantity. The mathematical quantity used to specify the amount of measurable information is the reciprocal of the variance, or the invariance, of the estimate.

The class of estimates which, as the sample is increased without limit, tend to be distributed about their limiting value (their mathematical expectation) in the normal distribution is the one appropriate to the theory of large samples. The amount of information afforded by an estimate normally distributed with variance V is $1/V$, the invariance of that normal distribution. In the normal case, the variance decreases with increasing size of sample, n, always ultimately in inverse proportion to n.

The criterion of efficiency, noted above, is that the limiting value of nV, where V is the variance of the estimate, shall be as small as possible. Fisher (Ref. 11) proved mathematically that the limiting value of $1/nV$ cannot exceed a quantity i, the amount of information provided by each observation the value of which is independent of the method of estimation. It was shown that the reciprocal of the variance, or the invariance of the estimate, cannot exceed the amount of information in the sample. Thus:

$$\frac{1}{V} \leqq ni = I \tag{6.01}$$

This conclusion is dependent on proof that for certain estimates the limiting value of

$$\frac{1}{nV} = i \tag{6.02}$$

The Maximum Likelihood Estimate. The instrument supplied by Fisher for obtaining the estimates necessary for the limiting value (6.02) to hold is the *method of maximum likelihood.* By this method, estimates of the parameters are obtained which maximize the likelihood function and have the smallest limiting vâriance. The limiting value of the· sampling variance of the maximum likelihood variance in large samples was proved to be

$$\frac{1}{nV} = i$$

We may state here that the probability of occurrence of a sample is expressible as a function of the unknown parameters, and the likelihood is defined as a function of these parameters proportional to this probability. Thus, the method of maximum likelihood gives as estimates those values which maximize the probability that the totality of observations should be that observed if the hypothesis which specifies the parameters of the population sample is true.

In large samples the maximum likelihood estimate has the smallest variance in comparison with any other statistic which is in the limit normally distributed. If the comparisons were restricted to statistics which in the limit are normally distributed, the utility of this method of estimation would be greatly limited. However, a stronger property than efficiency is possessed by the maximum likelihood estimate. This property exists when estimates may be made which contain within themselves the whole of the information available for finite samples. This is the property of *sufficiency.* Where sufficient statistics exist, all the available information is contained in the maximum likelihood estimate. In random samples from a normal population, the mean and the standard deviation—the only two characteristics necessary to specify this population—are sufficient statistics. It is this fact that gives the great simplicity to the problems falling within the theory of errors. Thus, in much experimental work it is necessary to be concerned only with the precision of the sum, or mean, of the observational values and with the estimation of this precision from the sum of squares calculated from the data. These two quantities contain all the information provided by the data with respect to the mean and variance of the hypothetical normal model. In cases where no sufficient statistic exists, Fisher has shown how the information in the sample may be recovered by using as ancillary the configuration of the sample. The configuration serves to indicate the precision of the estimate made, although it gives no information about the value of the parameter itself.

In experiments where the variance of the population is not known, it must be estimated from the data. Such an estimate is itself subject to error. For this error, exact allowance is made in the distribution of t when we test the significance of the deviation of the observed value from a hypothetical value specified by hypothesis. In such cases it would be inexact to assume that the amount of information provided by the experimental results with respect to the true value under estimation would be given by $1/s^2$, the reciprocal of the sampling variance. In determining the absolute precision of the experimental result, not only the estimate, s^2, derived from the data but also the number of degrees of freedom used in the estimate need to be taken into account. In this case it has been shown (Ref. 11, page 249) that the amount of information provided by an observed value, x, relative to the unknown mean population value, μ, is given by

$$\frac{n+1}{(n+3)s^2} \tag{6.03}$$

where n is the number of degrees of freedom.

Other Methods of Estimation. The most important general method of estimation so far discovered, at least from the theoretical standpoint, is the method of maximum likelihood. It will be frequently encountered in later discussions. There are other methods of estimation which should be considered. Under certain conditions all methods may yield similar results.

The oldest general method of forming estimates of the parameters of a distribution from sample values is the *method of moments* introduced by Karl Pearson, in which sample moments are equated to the corresponding moments of the distribution which are functions of the unknown parameters. As many moments as there are parameters requiring estimation are taken into account. The obtained equations with reference to the parameters are solved to give the estimates of the parameters. The fitting of the normal curve to a series of observations illustrates the process of the method of moments. The moment coefficients often involve relatively simple calculations in practice, but their efficiency decreases when the variations among the observations depart widely from normality.

The criterion of testing the closeness of an estimate in terms of a minimum standard deviation of its sampling distribution has been considered. Likewise, the criterion of testing closeness of fit of the estimates to certain parameters by a minimum χ^2-value has been used. Both these criteria are satisfied by the method of maximum likelihood in dealing with large samples.

The original use of the χ^2-test by Pearson (Ref. 27) was in the case of a completely specified hypothetical distribution. In this case it was established that χ^2, under the assumption that the hypothesis is true, is distributed in repeated sampling in a χ^2-distribution with $r - 1$ degrees of freedom (r is the number of groups into which the sample values have been classified). Most often in practice, the hypothetical distribution contains one or more unknown parameters. In these cases certain modifications were necessary in finding the limiting distribution of χ^2. Fisher (Refs. 9 and 4) showed that, for certain important methods of estimation, the modification could be made by reducing the number of degrees of freedom of the limiting distribution of χ^2 by one for each estimated parameter.

The method of estimation yielding a minimum χ^2 value is known as the χ^2 *minimum method of estimation.* In practice, the method often leads to difficult solutions, so that certain modifications have resulted in what is known as the *modified* χ^2 *minimum method* (Ref. 2, page 426). In certain cases this method is identical with the maximum likelihood method. In the case of fitting certain distributions, for example the binomial and Poisson distribution, and the normal distribution, the two methods give the same results. The method of maximum likelihood, however, can be extended to problems more general in nature.

A method of estimation developed by Markoff (Refs. 21 and 26) is based on the principle of unbiased estimates. Markoff has shown in various cases how to construct linear forms in the observational data which give estimates of certain unknown parameters that have no bias and the variances of which have the smallest possible value. The process of obtaining the best unbiased estimate of the population variance, σ^2, is based on this principle, for example, $s^2 = \dfrac{\Sigma(X - \bar{X})^2}{n - 1}$

Point Estimation and Its Limitations. The procedures of estimation just discussed may be called *estimation by a point.* A single value is given as the "best" estimate of the true or population value. Such a procedure does not provide a basis for specifying the degree of confidence one may place in such an estimate. It is known, of course, from sampling theory that the estimate made is not likely to be exactly equal to the population value. With large homogeneous samples the discrepancy is small, but with small samples the discrepancy may be considerable. Point estimation does not take directly into account the size of the sample which supplies the unique estimate. Because of these limitations in the method of point estimation, estimation by intervals seems to be increasing in use.

There are, of course, many occasions when a single value estimate is needed, particularly for certain subsequent statistical analyses. In the case of interval estimation, the single estimate is wanted as material for a subsequent process of estimation.

Estimation by Interval. We cannot tell from any sample estimate whether it is too great or too small. For this purpose further samples from the same population would be needed. It seems obvious therefore, that what is required is an interval of some kind which may be expected to include or cover the true population value in a specified number of cases. From the sample value and other ancillary information, we can calculate the point values of the upper and lower limits of the interval and then proceed to state that this interval will include or cover the population value. From sampling theory we can calculate the number of times in repeated sampling that the statement would be correct. Thus, the proportion of cases in which the statement may be assumed to be correct provides a measure of the confidence to be ascribed to our statement.

Fiducial Limits. R. A. Fisher (Refs. 10 and 3) first introduced the method of estimation based on the concepts of *fiducial probability* and *fiducial limits.* The basic ideas underlying Fisher's theory may be presented as follows.

Observations in the experimental or observational sciences are concrete and specific occurrences. They are now freely applied as a basis for probability statements about parameters whose exact values are unknown except for the information available in the observations. The kind of reasoning employed here comes from tests of significance, and the probability statements are designated as statements of fiducial probability, in order to distinguish such statements from those about "inverse probability." Fisher (Ref. 12) has indicated the fundamental random-variable relation which connects sample and population. The essential step in establishing this relation is in the following assertion: Irrespective of the character of the sample, the probability that the population parameter shall fall in any range is derived from the known probability, P, which is defined as the function of the variable, and from the test or the pivotal quantity in the test of significance. The assertion requires only that the unknown parameter value shall fall in the range corresponding to these known quantities. In this sense is to be interpreted the somewhat paradoxical statement that a sample with *known* characteristics is a *random* sample of an unknown population.

The properties of variable statistics are derived from observations which are defined as random variables involving parameters upon which their distribution functions are dependent. These properties are used to establish the connections between the probability distribution of the random variable and the distribution of the statistic used as the pivotal quantity in the test of significance. The statistic used as the pivotal quantity is functionally independent of the population from which the sample is drawn. This connection, once established, gives meaning to the practical situation where the statistics are observable but the parameters are unknown.

An illustration (Ref. 12) serves to show the application of this process of reasoning or form of fiducial argument. Following it, one may go from forms of statements embodying observations as random variables to forms of statements embodying observations as fixed data. In the former, the distribution functions include certain fixed but unknown parameters; in the latter, the frequency distributions are derived for the unknown parameters considered as random variables.

Let ξ be the median of a distribution concerning which the only thing known is that its probability integral is continuous. Take the case where $n = 2$, that is, where X_1 and X_2 are two observational values of the variable X. For any given value of ξ, the facts are that the three probabilities—that X_1 and X_2 (a) should both exceed the median, (b) should lie on either side of it, (c) should both be less than it—must occur in the frequency ratio 1:2:1. If r stands for the number of observations less than the median, then r becomes a pivotal quantity involving both the unknown parameter and the observations with a sampling distribution independent of the parameter; that is, r takes the values 0, 1, and 2 with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. This leads to the fiducial argument from the two given observations, now considered as fixed parameters, that the probability is .25 that ξ is less than both X_1 and X_2; .50 that ξ lies between X_1 and X_2; and .25 that ξ exceeds both X_1 and X_2. This reasoning thus leads to a frequency distribution of ξ, now considered as a random variable.

For a sample of any size, n, the following quantity expresses the probability that the median shall exceed r of the observations and be less than $n - r$:

$$\frac{n!}{r!(n-r)!}2^{-n} \tag{6.04}$$

Confidence Intervals. The complete theory underlying the method of interval estimation developed by Neyman (Ref. 25) cannot be presented here. However, the definition and use of the two concepts of confidence intervals and of confidence coefficients are presented briefly.

Consider a sample of n random variables $X_1, X_2, \ldots, X_n$, the n observational values. Denote by E the set of values of the X variables. This set can be represented by a point, called the *sample point* E in an n-dimensional space, the rectangular coordinates of E being $X_1, X_2, \ldots, X_n$. Assume that the probability law of the sample $X_1, X_2, \ldots, X_n$, though known, is given in terms of two parameters θ_1 and θ_2, which are unknown. It is desired to make an estimate of one of the parameters, say θ_1.

The process of estimating θ_1 consists in constructing two functions of the observations, $\underline{\theta}(E)$ and $\bar{\theta}(E)$ and in estimating the parameter to be within the interval: $\delta(E) = [\underline{\theta}(E), \bar{\theta}(E)]$. It is important to point out certain properties of the functions $\underline{\theta}$ and θ. Since they are functions of the sample values, $X_1, X_2, \ldots, X_n$, they are both random variables and will vary from sample to sample as the sample point $X_1, X_2, \ldots, X_n$ varies. Since they are random variables, the probabilities of $\underline{\theta}$ and $\bar{\theta}$ lying within or without any specified limit may be considered.

Denote by θ_1^0 the true value of the parameter θ_1 in a particular problem. Then $\underline{\theta}(E)$ and $\bar{\theta}(E)$ should have this property: the probability that when θ_1^0 and θ_2 are the true values of the two parameters, $\underline{\theta}(E)$ is less than θ_1^0 and $\bar{\theta}(E)$ is greater than θ_1^0 and is equal to α; that is,

$$P[\underline{\theta}(E) < \theta_1^0 < \bar{\theta}(E) | \theta_1^0, \theta_2] = \alpha \tag{6.05}$$

The interval extending from $\underline{\theta}(E)$ to $\bar{\theta}(E)$ in (6.05) is called the *confidence interval* corresponding to the sample point E, and the value α (for example, 0.95, or 0.99 . . .), the *confidence coefficient.* What is required in (6.05) is a probability of a specified value, whatever the values of θ_1 and θ_2, calculable from the probability law depending on θ_1 and θ_2. Thus the functions $\underline{\theta}$ and $\bar{\theta}$ must satisfy (6.05), also identically for all possible values of θ_2.

The meaning of the confidence interval may be said to be this: Assume that a large number of samples are drawn randomly from a population obeying the specified elementary probability law. If in each case the statement is made that θ_1^0 is included in the interval $[\underline{\theta}(E), \bar{\theta}(E)]$, then the relative frequency of correct statements will be approximately equal to the confidence coefficient, α. For example, take $\alpha = 0.95$. If 100 samples are taken and 100 confidence intervals set up, it may be expected that 95 per cent of these intervals will include or cover the true value, say θ_1^0. It should be noted that this statement is not equivalent to the statement that the probability is 95 out of 100 that θ_1^0 lies between the limits $\underline{\theta}$ and $\bar{\theta}$. This discrepancy is explained by the fact that θ_1^0 is not a random variable but an unknown constant. Consequently, the probability of θ_1^0 falling within specified limits may be either zero or unity, depending on whether the actual value of θ_1 falls without or within the limits.

Further development of the theory (Ref. 24) indicates that there exists an infinite number of confidence intervals for a given confidence coefficient. Hence, some principle is needed as a basis for choosing from among them. One principle is to select the shortest system of intervals. Shortest confidence intervals, however, exist to a considerable extent only in exceptional cases. Other principles, such as unbiasedness, have been used; but even shortest unbiased confidence intervals exist in only a restricted class of cases. A third type of interval has been called the "short-unbiased" confidence interval. If there is more than one parameter, there is not often a confidence interval for one of the parameters which is independent of the other parameters. With more than one parameter the set of points constitutes a simple close region, if it exists, rather than a single interval as in the case of only one parameter. In the case of several parameters, new problems arise. But the description of the basic ideas has been given in the situation described above.

Fiducial versus Confidence Intervals. It appears that Fisher's theory of fiducial probability and Neyman's theory of confidence intervals are closely related and that in a number of practical cases they may lead to the same form of procedure. The authors, however, indicate a disagreement in the logical foundations as well in certain practical applications. Neyman (Ref. 23) has attempted to develop a general procedure which will supply rules for setting up from observational data an interval that will cover the unknown parameter with a given probability. Fisher (Ref. 7) indicates that a unique probability measure associated with a particular interval is needed. This measure is defined as a fiducial probability. An essential point of agreement is in the interpretation that the probability of, say, 0.95 is not the probability that the parameter estimated lies between any fixed limits but, rather, that a variable statement about this parameter formulated in accordance with a specified rule will be correct. Fisher expresses it by stating that there is a fiducial probability of 95 per cent of the unknown parameter's lying within the specified fiducial limits. According to Neyman, the statement would be made that the specified interval will cover the true value and that we know that the statement will be correct 95 times out of 100.

Fisher (1935) has emphasized that a fiducial statement can be made only in terms of the estimate if the estimate of the unknown parameter has the property of sufficiency, because only in this case does the estimate elicit the whole of the available information. Neyman's confidence intervals are apparently of more general applicability. When an estimate is sufficient, both the fiducial limits and the limits of Neyman's shortest confidence interval or of his short unbiased confidence interval depend on this property of sufficiency. The interval would not, however, always be the same in the two cases because of the use by Neyman of an additional principle in the determination of his intervals.

It would appear, however, that the two procedures would be interchangeable in at least the first two examples that follow.

PROBLEMS OF INTERVAL ESTIMATION

Problem VI.1. Estimation of the population mean. The first problem consists in estimating μ, the mean of a normal population of known variance σ^2, given a sample mean $\bar{X}$ based on n items.

From our study of sampling theory, we know that the means of random samples from a normal population, for example, the $\bar{X}$'s, are normally distributed about μ with a standard deviation (called the *standard error of the mean*, $\sigma_{\bar{x}}$) equal to $\sigma/\sqrt{n}$. Hence, we know the proportion of sample means which will lie within the interval: $\mu \pm$ some multiple of $\sigma_{\bar{X}}$. The confidence interval may be written as

$$\bar{X} - y_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + y_\alpha \frac{\sigma}{\sqrt{n}} \tag{6.06}$$

where y_α is the value of

$$y = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

for a given confidence coefficient α, which can be read from a normal probability integral table. If $\alpha = 0.99$, then $y_\alpha = 2.576$. If $\alpha = 0.95$, then $y_\alpha = 1.96$ no matter what n is. For example, we find that 99 per cent of the sample means will fall within the interval $\mu \pm 2.576\sigma_{\bar{X}}$, and 95 per cent within the interval $\mu \pm 1.96\sigma_{\bar{X}}$. On the basis of sampling theory, if in repeated sampling we take the interval extending from a lower limit of $\bar{X} - 2.576\sigma_{\bar{X}}$ to an upper limit of $\bar{X} + 2.576\sigma_{\bar{X}}$, then this interval will cover the population mean, μ, in 99 per cent of cases.

We may take as a practical illustration the 100 samples of 5 items each drawn from the population with $\mu = 30$, $\sigma = 10$. For samples of 5 numbers $\sigma_{\bar{X}}$ will be

$$\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{5}} = 4.472$$

Using a 95 per cent confidence coefficient, we take the intervals extending from $\bar{X} - (1.96)(4.472)$ to $\bar{X} + (1.96)(4.472)$, or from $\bar{X} - 8.77$ to

TABLE 31

CONFIDENCE INTERVALS FOR THE MEANS OF 100 RANDOM SAMPLES OF SIZE 5, USING A CONFIDENCE COEFFICIENT OF 95 PER CENT

(Population $\mu = 30$; $\sigma = 10$)

14.23–31.77	29.83–47.37	12.83–30.37	25.03–42.57	24.83–42.37
28.03–45.57	18.23–35.77	19.83–37.37	20.23–37.77	15.03–32.57
18.63–36.17	19.63–37.17	21.03–38.57	23.43–40.97	19.23–36.77
24.03–41.57	18.83–36.37	22.03–39.57	31.83–49.37	23.23–40.77
30.23–47.77	23.43–40.97	18.43–35.97	21.03–38.57	15.03–32.57
23.43–40.97	25.83–43.37	24.43–41.97	13.83–31.37	27.03–44.57
23.03–40.57	23.63–41.17	21.83–39.37	20.03–37.57	24.03–41.57
16.83–34.37	17.83–35.37	28.83–46.37	22.63–40.17	16.83–34.37
14.63–32.17	22.83–40.77	23.23–40.77	23.83–41.37	23.43–40.97
23.03–40.57	23.03–40.57	19.43–36.97	17.23–34.77	17.23–34.77
22.43–39.97	21.03–38.57	26.03–43.57	20.83–38.37	14.43–31.97
26.23–43.77	21.83–39.37	23.23–40.77	14.83–32.37	19.63–37.17
20.43–37.97	22.43–39.97	25.83–43.37	23.03–40.57	22.23–39.77
19.03–36.57	17.63–35.17	29.23–46.77	21.63–39.17	18.43–35.97
13.03–30.57	18.23–35.77	17.63–35.17	19.43–36.97	12.83–30.37
13.83–31.37	16.83–34.37	20.43–37.97	23.63–41.17	29.03–46.57
29.03–46.57	19.23–36.77	25.03–42.57	13.63–31.17	15.03–32.57
13.03–30.57	15.23–32.77	15.83–33.37	30.03–47.57	20.63–38.17
15.83–33.37	22.63–40.17	15.43–32.97	28.83–46.37	20.43–37.97
10.63–28.17	26.23–43.77	29.63–47.17	22.63–40.17	23.83–41.37

$\bar{X} + 8.77$. We calculated these intervals for the 100 sample means given in Table 3, page 34. They are recorded in Table 31. It is noted that only one of the 100 intervals calculated, namely, 10.63 – 28.17, does not include the population mean 30.

The sampling experiment was repeated by taking random samples of size 50 instead of 5. The means of 100 samples of 50 items each were calculated. Again, the intervals were set up by using a confidence coefficient of 95 per cent, which in this case extended from $\bar{X} - (1.96)(1.4142)$ to $\bar{X} + (1.96)(1.4142)$, or from $\bar{X} - 2.77$ to $\bar{X} + 2.77$. We found that the population mean 30 was covered in 97 of the 100 cases.

An extension of the sampling experiment was made to obtain the means of 100 samples of 100 items each. The confidence intervals with a confidence coefficient of 95 per cent were calculated again, given by the limits $\bar{X} - 1.96$ and $\bar{X} + 1.96$. We noted that the population mean 30 was covered in 96 of the 100 cases.

In all three of the sampling experiments, therefore, there was a close agreement between theory and observation. We noted also that the confidence intervals become shorter as the size of the sample is increased. Therefore, the larger the sample, the more accurately can the true or population value be estimated.

Problem VI.2. Estimation of the population mean of a normal population of unknown variance. Nearly always, in experimental work, neither the mean nor standard deviation of the population from which we are sampling is known. In estimating the population mean in such cases, we have to use the mean and standard deviation of the sample and the distribution of *t*. We shall calculate the fiducial values according to Fisher (Ref. 5, pp. 195–198).

A fundamental principle in the use of the *t*-distribution for the solution of this problem is: If an estimate of a parameter is normally distributed with a variance which can be estimated from the sample and the distribution of which is independent of the estimate of the parameter, then fiducial limits can be calculated from "Student's" ratio.

The following are the characteristics of *t* which give it its unique utility for the solution of this type of problem:

(a) The distribution of *t* is known with exactitude, without any supplementary assumptions or approximations.

(b) *t* is given by the single unknown parameter, μ, and by observable statistics only.

(c) The statistics involved in the quantity *t* are sufficient.

The quantity *t* is expressed by:

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{\sqrt{n}\,(\bar{X} - \mu)}{s} \qquad (6.07)$$

where $$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

It is noted that since all terms on the right-hand side of (6.07) except μ are observable, the fiducial values of μ are determinable when values of t appropriate to any chosen level of significance, ϵ, have been chosen. Furthermore, $\bar{X}$ and s^2 are independently distributed; and the two quantities, the sum and the sum of squares, calculated from the data are sufficient statistics, since they contain all the relevant information concerning the mean and variance of the hypothetical normal curve. Therefore, we may write

$$\mu = \bar{X} \pm t_\epsilon \sqrt{\frac{s^2}{n}} \tag{6.08}$$

as the corresponding fiducial limits for the value of μ. With respect to μ, it may then be said that the fiducial probability is $(1 - \epsilon)$ that it will lie within these fiducial limits.

As a practical illustration, we may set up the fiducial limits of the true mean difference based on the data from the controlled experiment given in Problem V.6, page 75, in which the null hypothesis was rejected at the 5 per cent level.

The following quantities were obtained:

$$\begin{aligned} \bar{X} &= 9.28 \\ n &= 25 \\ \Sigma(X - X)^2 &= 6809.04 \\ s^2 &= \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{6809.04}{24} = 283.71 \end{aligned}$$

We wish to set up fiducial limits with a fiducial probability of 95. Accordingly,

$$\begin{aligned} t_\epsilon &= t_{.05} \\ \pm t_{.05} &= \pm 2.064 \qquad \text{(for } n - 1 = 24\text{)} \end{aligned}$$

and

$$\left.\begin{aligned} \mu &= \bar{X} \pm t \sqrt{\frac{s^2}{n}} \\ &= 9.28 \pm (2.064)(3.368) \\ &= 2.33 \text{ or } 16.23 \end{aligned}\right] \tag{6.09}$$

With respect to the population mean μ, it may be said that it has a fiducial probability of 2.5 per cent of being less than 2.33 or of being greater than 16.23, and, in the same sense, a probability of 95 per cent of lying within these fiducial limits.

Problem VI.3. Estimation of the population variance from the sample value. If the experimenter is interested in determining whether the variance or the standard deviation of a normal distribution could exceed a given value or could lie in a given range, a test of significance is needed for which the pivotal quantity should possess the following characteristics:

(a) Its exact sampling distribution must be known.
(b) It must be expressible in terms of the unknown variance, ϕ, of the distribution sampled, together with known statistics only.
(c) The statistics involved in the expression of the quantity must be sufficient.

It is known that $\frac{(n-1)s^2}{\phi}$ is distributed as is χ^2 for $n - 1$ degrees of freedom. That is, if $\frac{\chi^2}{n-1}$ is the ratio of the estimate of the population variance as obtained from the sample for $n - 1$ degrees of freedom to the true variance, ϕ, then χ^2 is distributed, independently of the population. mean and variance, in a distribution determinable from the number of degrees of freedom (Ref. 6).

The upper and lower hundred ϵ per cent fiducial limits of ϕ can be obtained from tables of the χ^2-distribution. If the two critical values of χ^2 are represented by χ_1^2 and χ_2^2, the fiducial range of ϕ will be the interval

$$\left[\frac{(n-1)s^2}{\chi_1^2}, \frac{(n-1)s^2}{\chi_2^2}\right]$$

As our practical illustration, we set up the fiducial limits of the variance of the distribution based on the data in Problem V.6. If we take $\epsilon = .05$ as the probable lower limit of the value of ϕ for $n - 1 = 24$, χ^2 is less than 36.415 in only 5 per cent of trials (see Table III, Appendix).

Substituting this value of χ^2 in the equation

$$\left.\begin{aligned} \chi^2 &= \frac{\Sigma(X - \bar{X})^2}{\phi} \\ &= \frac{6809.04}{\phi} \end{aligned}\right] \tag{6.10}$$

We have

$$\begin{aligned} \phi &= \frac{6809.04}{36.415} \\ &= 186.98 \end{aligned}$$

Similarly, the probable upper limit to the value of ϕ is obtainable by first noting that χ^2 for $n - 1$ degrees of freedom exceeds the value 13.848 in only 5 per cent of trials. Substituting this value for χ^2 in Equation (6.10),

$$\chi^2 = \frac{6809.04}{\phi}$$

we get

$$\begin{aligned} \phi &= \frac{6809.04}{13.848} \\ &= 491.70 \end{aligned}$$

We may say, then, that the fiducial probability is 5 per cent that the variance should exceed 491.7 or be less than 186.98, and, in the same sense, a fiducial probability of 90 per cent of the variance lying within these fiducial limits. If a linear measure of variation is wanted, the corresponding fiducial limits for the population standard deviation are 22.17 and 13.64, respectively.

Problem VI.4. Estimation of an individual's true score from his obtained score on a test. We assume that the scores an individual would obtain on a very large number of equivalent tests are distributed in a normal manner about his true score with a standard deviation equal to the standard error of an individual score, $\sigma_x = s\sqrt{1 - r}$, where s is the standard deviation of the distribution of scores and r is the reliability coefficient of the test. The upper and lower limits of the confidence interval of his true score, ξ, are given by

$$X \pm Y_\alpha(s\sqrt{1 - r}) \tag{6.11}$$

where Y_α is the value of $y = (X - \xi)/\sigma_x$ for a given confidence coefficient, α, which is read from the normal probability integral table; X is the obtained score; and σ_x is the standard error of X.

As an illustration, let us set up the confidence interval for the true score of a pupil who receives an I.Q. rating of 105 on a particular intelligence test on which the standard error of an individual score is 4 I.Q. points. Using a confidence coefficient of 98 per cent, the upper and lower limits of the confidence interval are, respectively, $105 + (2.326)(4) = 114.3$ and $105 - (2.326)(4) = 95.7$. We then state that the interval (95.7, 114.3) will cover the true I.Q. score of this individual, and we know that our statement concerning the true score, ξ, will be correct in 98 per cent of such cases.

Problem VI.5. Estimation of the confidence interval for the population median in samples from any continuous population. We have considered the sampling distributions of certain statistics calculated from random samples involving only one of the unknown parameters specifying a parent population of known form. The method of interval estimation was used to set up in terms of the observations at any level the confidence interval for the unknown population parameter.

Thompson (Ref. 30) and Savur (Ref. 29) independently obtained the confidence interval for the median without reference to the form of parent population. Cases arise in which the population form is unknown or, as in small samples, in which it is not easy to test an assumption of normality. Here the interval estimation of the median as a measure of location is especially useful. Nair (Ref. 22) used the results of Thompson, restricted to continuous populations, to construct a table of confidence intervals for the median, the use of which makes the problem of estimation extremely simple.

In a random sample of n observations $X_1, X_2, \ldots, X_k, \ldots, X_n$ arranged in ascending order of magnitude, if P_k is the probability integral of X_k, then

$$P(X < X_k) = P(P < P_k) = \int_P^1 P^{k-1}(1 - P)^{n-k}dp = I_{1-p}(n - k + 1,k) \tag{6.12}$$

where $I_X(P,q)$ is the function tabulated in the Incomplete Beta Function Table. By definition, the probability integral corresponding to the median, M, is $\frac{1}{2}$. Therefore,

$$P(M < X_k) = P(\tfrac{1}{2} < P_k) = I_{0.5}(n - k + 1,k)$$

Also, $$P(M < X_k) = P(M > X_{n-k+1})$$

Hence, $$P(X_k < M < X_{n-k+1}) = 1 - 2I_{0.5}(n - k + 1, k) \tag{6.13}$$

which is the confidence interval of the population median. It states that the unknown population median will lie in the interval extending from the kth to the $(n - k + 1)$th observation in 100 $[1 - 2I_{0.5}(n - k + 1, k)]$ per cent of the cases.

With the aid of the Incomplete Beta Function Tables, Nair (Ref. 22) prepared the Table of Confidence Intervals for the Median for values of n from 6 through 81, for confidence coefficients of 0.95 and 0.99. This table consists in finding k such that, given n,

$$I_{0.5}(n - k + 1, k) = 0.025 \text{ or } 0.005$$

Since k can have only integral values, the confidence coefficient cannot be fixed exactly at 0.95 or 0.99 for all values of n. Values of k are taken, which bring the confidence coefficient

$$I - 2I_{0.5}(n - k + 1, k)$$

nearest to (and greater than) the conventional values of 0.95 or 0.99.

For values of n larger than 81, Nair (Ref. 22) suggests the use of the normal curve as an approximation where x, the relative deviate, is given by

$$x = \frac{\left(\frac{n}{2}\right) - k}{\left(\frac{\sqrt{n}}{2}\right)} = \frac{n - 2k}{\sqrt{n}} \tag{6.14}$$

For a given confidence coefficient, such as 0.95 or 0.99, the corresponding value of x can be obtained from the Normal table, and the value of k can be determined from the relation

$$k = \frac{n - x\sqrt{n}}{2} \tag{6.15}$$

As an illustration of the use of Nair's Tables, we shall set up the confidence interval for the population median, M, using the sample data given in Problem V.6, page 75. The sample median, Md, of the individual pair differences is 10; $n = 25$.

We enter Nair's Table given in (Ref. 22) with $n = 25$. The arguments and values for $n = 25$ given in the table are as follows:

Confidence coefficient ≧ 0.95				Confidence coefficient ≧ 0.99			
n	k	$n - k + 1$	$I_{0.5}(n - k + 1, k)$	k	$n - k + 1$	$I_{0.5}(n - k + 1, k)$	n
25	8	18	.0216	6	20	.0020	25

For a sample of 25 observations, we can say that, with a confidence coefficient of 95.68 per cent [100 − 2(.0216)], the population median M will lie between the eighth and eighteenth ranked observations, that is, between 17 and 3. We can also say that, with a confidence coefficient of 99.6 per cent, the population median, M, will lie between the sixth and twentieth ranked observation, that is, between 24 and −6.

Problem VI.6. Setting up the confidence interval on a population difference from a given sample difference in percentages. If percentages are obtained within a sample, that is, the percentages of "yes" and "no" answers to a given question, the problem arises of how to get confidence limits on a population difference, d, for a given sample difference, $\bar{d}$. Wilks (Ref. 32) gives the 99 per cent sampling limits of $\bar{d}$ as

$$d \pm \frac{2.58}{\sqrt{n}} \sqrt{100(P_1 + P_2) - d^2} \tag{6.16}$$

That is, in drawing repeatedly random samples of size n from a population in which the "yes" and "no" percentages are P_1 and P_2, respectively, approximately 99 per cent of the samples have a difference $\bar{d}$ which lies between these two limits. In practice, sample values $\bar{d}$, $\bar{P}_1$, and $\bar{P}_2$ are substituted for the unavailable d, P_1, and P_2. This procedure may be satisfactory for practical purposes.

Wilks gives the quantity $258/\sqrt{n}$ as a simple, conservative critical value of he sample difference $\bar{d}$. If $+\bar{d}$ is larger than $258/\sqrt{n}$, the probability is at least 0.99 that d, the population difference, would be included between two positive confidence limits. The more common interpretation would be that at the 1 per cent level of significance a true difference d between the "yes" and "no" percentages exists in the population.

As an illustrative problem, let us test the significance of the difference between the two proportions in a random sample of 77 male graduates of a teachers' college, 22 of whom remained in the teaching profession and 55 of whom left it within 10 years after graduation. The approximate test of significance given by the pivotal quantity $258/\sqrt{n}$ shows that d, or 42.8 per cent, $> 258/\sqrt{77} > 29.4$ and hence significant at the 1 per cent level. The 99 per cent confidence interval is given by substituting the sample values for d, P_1, and P_2 in (6.16). Thus:

$$d \pm \frac{2.58}{\sqrt{77}} \sqrt{100(71.4 + 28.6) - (42.8)^2} = 42.8 \pm 26.5$$

or the confidence interval of the population difference, d, with a confidence coefficient of 99 per cent is (16.3, 69.3).

Problem VI.7. Setting up a confidence interval of a population difference from the difference between two sample percentages. The problem of comparing two percentages in different random samples differs from that in Problem VI.6 in that in the latter there is a negative correlation between the percentages of "yes" and "no" answers. No correlation exists in the percentages in the two different samples. Wilks (Ref. 32) gives 99 per cent sampling limits of $\bar{d}$ as

$$d \pm 2.58 \sqrt{\frac{P_1(100 - P_1)}{n_1} + \frac{P_2(100 - P_2)}{n_2}} \tag{6.17}$$

and the corresponding conservative critical limit for $\bar{d}$ as

$$129 \sqrt{\frac{(n_1 + n_2)}{n_1 n_2}} \tag{6.18}$$

If instead of calculating $\bar{d}$, the difference between the percentages P_1 and P_2, we first transform the percentages to the inverse sine function (see page 164), then

$$\bar{d}' = 100 \left(\sin^{-1} \sqrt{\frac{\bar{P}_1}{100}} - \sin^{-1} \sqrt{\frac{\bar{P}_2}{100}} \right) \tag{6.19}$$

Then $129 \sqrt{\frac{(n_1 + n_2)}{n_1 n_2}}$ is, to a close approximation at the 1 per cent level, an exact critical limit of $\bar{d}'$.

As an illustration, we shall set up the 99 per cent confidence interval for the population difference from the two samples of percentages of color-blindness in the two sexes of the Caucasoid population (see Problem V.8, page 80).

From the data,

$$\bar{P}_1 = 8.4, \qquad \bar{P}_2 = 1.3, \qquad \bar{d} = 7.1$$

The 99 per cent confidence interval is obtained by substituting the sample values P_1, $\bar{P}_2$, and $\bar{d}$ in (6.17):

$$d \pm 2.58 \sqrt{\frac{(8.4)(91.6)}{793} + \frac{(1.3)(98.7)}{232}} = 7.1 \pm 3.18$$

Therefore, the 99 per cent confidence interval of the population difference is (3.9, 10.3).

Again using Formula (6.19), we get[1]

$$\begin{aligned} \bar{d}' &= 100(\sin^{-1}\sqrt{.084} - \sin^{-1}\sqrt{.013}) \\ &= 100(16.8 - 6.4) = 1040 \end{aligned}$$

and

$$129\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 129\sqrt{\frac{1025}{183{,}976}} = 9.6$$

Since 1040 > 9.5, the probability is at least 0.99 that a genuine difference between the percentages exists in the population.

Problem VI.8. Setting up the confidence limits for an individual estimated score. In problems of estimating or predicting a measure of a characteristic from a knowledge of one or more other characteristics, the predicted values are subject to error. Here we can use the confidence interval to show the accuracy of individual estimates and the confidence that may be placed in the statements made about individuals. We shall take the case of simple regression, that is, the prediction of one characteristic from a knowledge of another.[2] The data are from Problem V.13, page 88, and we shall set up the confidence interval for each of the individual's estimated score from the regression equation, using a confidence coefficient of 98 per cent. The basic calculations are given in Table 32.

The standard error of the estimate Y_E for a particular value of X, say X_0, is given by

$$s_{Y_E} = \left\{\frac{s_Y^2(1 - r^2)}{N - 2}\left[1 + \frac{(X_0 - \bar{X})^2}{s_X^2}\right]\right\}^{\frac{1}{2}} \tag{6.20}$$

where s_{Y_E} denotes the standard error of Y_E, N is the number of pairs of observations, and the other quantities have their customary meanings.

From the formula, it is noted that the errors of the estimates of Y increase as the quantity X_0 departs from the mean of the X-distribution; also that as the values of r and s_X become larger, the smaller become the errors of estimation, other factors being equal.

From Problem V.13, we record the following values:

$$\begin{aligned} Y_E &= .9873X - 0.68 \\ s_X^2 &= 157.30 \\ s_Y^2 &= 172.59 \\ r^2 &= 0.8885 \\ \bar{X} &= 53.24 \\ N &= 25 \end{aligned}$$

[1] Transformation obtained from Fisher and Yates's Table XII, page 56 of Ref. 13 in Chapter VII.

[2] For the multivariate case see page 343.

TABLE 32

Standard Errors of Estimated Values of Y for Different Values of X_0 with Corresponding 98 per Cent Confidence Intervals

Individual	X_0	Y	Y_E	s_{Y_E}	$t_{.02}s_{Y_E}$	Interval	Independent variables arranged in descending order of magnitude		
							X_0	s_{Y_E}	$t_{.02}s_{Y_E}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	46	52	44.74	1.056	2.64	42.10–47.38	73	1.707	4.27
2	38	38	36.84	1.439	3.60	33.24–40.44	73	1.707	4.27
3	64	63	62.51	1.205	3.01	59.50–65.52	67	1.357	3.39
4	73	65	71.39	1.707	4.27	67.12–75.66	66	1.305	3.26
5	61	58	59.55	1.075	2.69	56.86–62.24	66	1.305	3.26
6	34	33	32.89	1.675	4.19	28.70–37.08	64	1.205	3.01
7	57	49	55.60	0.955	2.39	53.21–57.99	64	1.205	3.01
8	66	63	64.48	1.305	3.26	61.22–67.74	61	1.075	2.69
9	25	24	24.00	2.255	5.64	18.36–29.64	61	1.075	2.69
10	30	26	28.94	1.926	4.81	24.13–33.75	59	1.006	2.51
11	45	33	43.75	1.094	2.73	41.02–46.48	57	0.955	2.39
12	73	71	71.39	1.707	4.27	67.12–75.66	55	0.923	2.31
13	45	43	43.75	1.094	2.73	41.02–46.48	55	0.923	2.31
14	55	63	53.62	0.923	2.31	51.31–55.93	52	0.919	2.30
15	66	70	64.48	1.305	3.26	61.22–67.74	51	0.929	2.32
16	49	46	47.70	0.965	2.41	45.29–50.11	50	0.944	2.36
17	64	65	62.51	1.205	3.01	59.50–65.52	49	0.965	2.41
18	45	46	43.75	1.094	2.73	41.02–46.48	46	1.056	2.64
19	61	62	59.55	1.075	2.69	56.86–62.24	45	1.094	2.73
20	52	46	50.66	0.919	2.30	48.36–52.96	45	1.094	2.73
21	67	68	65.47	1.357	3.39	62.08–68.86	45	1.094	2.73
22	59	53	57.57	1.006	2.51	55.06–60.08	38	1.439	3.60
23	55	55	53.62	0.923	2.31	51.31–55.93	34	1.675	4.19
24	51	52	49.67	0.929	2.32	47.35–51.99	30	1.926	4.81
25	50	48	48.68	0.944	2.36	46.32–51.04	25	2.255	5.64

Substituting these values in Equation (6.20) and using the X_0 for each of the 25 individuals, we obtain the s_{Y_E} for each individual. These values are recorded in column (5), Table 32. Using the confidence coefficient of 98 per cent, we find from the t-table that the value of $t_{.02}$ for $n = N - 2 = 23$ is 2.5. Therefore, for any particular value of X_0 the upper and lower limits of the confidence interval will be $Y_E + 2.5s_{Y_E}$ and $Y_E - 2.5s_{Y_E}$. The values of $2.5s_{Y_E}$ are given in column (6), and the values for the confidence interval, in column (7).

In column (10) the values of $t_{.02}s_{Y_E}$ have been recorded for values of

X_0 [column (8)] arranged in descending order of magnitude. It is clear from column (9) that the errors of estimation increase considerably as the value of X_0 recedes from the mean of the distribution of X. Correspondingly, the confidence intervals widen and reflect the increase in the errors of estimation.

CONFIDENCE LIMITS AND TOLERANCE LIMITS

A distinction should be made between confidence limits and tolerance limits. The latter has proved to be a useful statistical concept in the quality control of manufacturing products and probably can be applied to other fields.

The problem in setting up tolerance limits is that of determining limits from the sample information which will include, on the average, a specified proportion of the universe or population between them (Ref. 33).

For an example, let $\bar{X}_n$ be the sample mean and $s^2 = \sum_{i}^{n} \frac{(X_i - \bar{X})^2}{(n-1)}$ the sample variance estimate in a sample of size n. The tolerance limits L_1' and L_2', which between them will include, on the average, a proportion α of the universe, are given by

$$\bar{X} \pm t_\alpha \sqrt{\frac{n+1}{n}} \cdot s \tag{6.21}$$

The value of t_α can be obtained from the table of the t-distribution, when the value of α has been specified, for example as 99 per cent, 95 per cent, or whatever, and $n - 1$ is the number of degrees of freedom. In contrast, the confidence limits $\bar{X} \pm t_\alpha s_{\bar{x}}$ may be said to include the population mean, μ, with a confidence coefficient of α.

THE METHOD OF MAXIMUM LIKELIHOOD

We shall illustrate the method of maximum likelihood for determining the best estimates of the population values by applying the method to the derivation of the estimates of the five parameters required to specify a normal correlation surface (Refs. 28 and 19). The five parameters are the means of the two normal distributions of the variates X and Y, say μ and ξ, respectively; the two standard deviations, σ_X and σ_Y; and ρ, the correlation between X and Y. It is assumed that the regression both ways (X on Y or Y on X) is linear and that the variables X and Y are normally distributed.

If the variables X and Y are normally distributed, then the probability distributions of X, Y, and XY, namely, $P(X)$, $P(Y)$, and $P(X,Y)$, are

$$P(X) = \frac{1}{\sigma_X \sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma_X^2}} \tag{6.22}$$

$$P(Y) = \frac{1}{\sigma_Y \sqrt{2\pi}} e^{-\frac{(Y-\xi)^2}{2\sigma_Y{}^2}} \tag{6.23}$$

$$P(X,Y) = \frac{1}{2\pi\sigma_X\sigma_Y \sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(X-\mu)^2}{\sigma_X{}^2} - \frac{2\rho(X-\mu)(Y-\xi)}{\sigma_X\sigma_Y} + \frac{(Y-\xi)^2}{\sigma_Y{}^2}\right]} \tag{6.24}$$

With N pairs of values, the simultaneous probability distribution of all the N values of X and Y is

$$P(X_1, \cdots, X_N, Y_1, \cdots, Y_N) = \left(\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\right)^N e^{-\frac{1}{2(1-\rho)^2}\Sigma[\ldots]}$$
$$\left[\frac{(X-\mu)^2}{\sigma_X^2} - 2\rho\frac{(X-\mu)(Y-\xi)}{\sigma_X\sigma_Y} + \frac{(Y-\xi)^2}{\sigma_Y^2}\right] \tag{6.25}$$

To obtain the maximum likelihood estimates, the process consists in taking the partial derivatives of the probability functions with respect to the parameters, setting the resulting equations equal to zero, and then solving the simultaneous linear equations for the parameters. Here, then, the procedure consists in taking the partial derivatives of $P(X_1, \ldots, X_N, Y_1, \ldots, Y_N)$ with respect to μ, ξ, σ_X, σ_Y, and ρ. It is convenient to work with the natural logarithms. Hence, for (6.25) we have

$$\left.\begin{aligned} \log_e P = -N\log_e 2\pi - N\log_e\sigma_X - N\log_e\sigma_Y - \frac{N}{2}\log_e(1-\rho^2) \\ -\frac{1}{2(1-\rho^2)}\sum\left[\frac{(X-\mu)^2}{\sigma_X^2} - \frac{2\rho(X-\mu)(Y-\xi)}{\sigma_X\sigma_Y} + \frac{(Y-\xi)^2}{\sigma_Y^2}\right] \end{aligned}\right] \tag{6.26}$$

Then log P is differentiated with respect to μ and the equation is set equal to zero, giving

$$\frac{\delta \log P}{\delta\mu} = 0 = \frac{2}{2(1-\rho^2)}\frac{\Sigma(X-\mu)}{\sigma_X^2} - \frac{2\rho}{2(1-\rho^2)}\frac{\Sigma(Y-\xi)}{\sigma_X\sigma_Y} \tag{6.27}$$

From which $$\sigma_Y\Sigma(X-\mu) = \rho\sigma_X\Sigma(Y-\xi) \tag{6.28}$$

Likewise, differentiating log P with respect to ξ, setting the derivative equal to zero, and reducing, we get

$$\sigma_X\Sigma(Y-\xi) = \rho\sigma_Y\Sigma(X-\mu) \tag{6.29}$$

Assuming $\rho \neq 1$, $\sigma_X \neq 0$, $\sigma_Y \neq 0$, we get by solving equations (6.28) and (6.29) the optimum estimates:

$$\mu = \frac{\Sigma X}{N} \equiv \bar{X} \tag{6.30}$$

$$\xi = \frac{\Sigma Y}{N} \equiv \bar{Y} \tag{6.31}$$

Similarly, we may differentiate log ρ partially with respect to σ_X, σ_Y, and ρ, respectively; set the equations equal to zero; solve; and substitute the values given for μ and ξ in (6.30) and (6.31); obtaining

$$\sigma_X = \sqrt{\frac{1}{N}\left[\sum X^2 - \frac{(\Sigma X)^2}{N}\right]} \equiv s_X \tag{6.32}$$

$$\sigma_Y = \sqrt{\frac{1}{N}\left[\sum Y^2 - \frac{(\Sigma Y)^2}{N}\right]} \equiv s_Y \tag{6.33}$$

$$\left.\begin{aligned} \rho &= \frac{\sum XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left[\sum X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\sum Y^2 - \frac{(\Sigma Y)^2}{N}\right]}} \equiv r \\ \text{or} \qquad \rho &= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \equiv r \end{aligned}\right] \tag{6.34}$$

Jackson and Ferguson (Ref. 19) have shown that the maximum likelihood estimate of ρ in the case of samples from a population specified by four parameters—$\sigma_X = \sigma_Y = \sigma$; μ; ξ; ρ—is

$$\rho = \frac{2\left[\sum XY - \frac{(\Sigma X)(\Sigma Y)}{N}\right]}{\left[\sum X^2 - \frac{(\Sigma X)^2}{N}\right] + \left[\sum Y^2 - \frac{(\Sigma Y)^2}{N}\right]} \tag{6.35}$$

This is the case in determining the reliability coefficient of a test by the test-retest and alternative equivalent forms methods.

In the same way, the maximum likelihood estimate of ρ obtained from samples of a population specified by three parameters—$\sigma_X = \sigma_Y = \sigma$; $\mu = \xi$; ρ—is

$$\rho = \frac{2\sum XY - \frac{(\Sigma X + \Sigma Y)^2}{2N}}{\sum X^2 + \sum Y^2 - \frac{(\Sigma X + \Sigma Y)^2}{2N}} \tag{6.36}$$

This is the case in determining the reliability coefficient of a test by the split-half method.

ESTIMATING THE RELIABILITY OF TESTS

The reliability of measurements is a fundamental tenet in all observational and experimental sciences. The problem of the reliability of instruments of measurement has, however, received the greatest consideration in psychology, education, and sociology.

The traditional method of determining the reliability of a test is through the use of the product-moment correlation coefficient. The term "reliability of a test" as introduced by Spearman in 1910 was defined as the (correlation) "coefficient between one half and the other half of several measurements of the same thing." (Ref. 19.)

Until recently, the only methods available for measuring the so-called "reliability of a test" were (1) the test-retest method—doing the same test twice; (2) obtaining the correlation between the scores on equivalent forms of the test; (3) the split-test method—consisting in obtaining the correlation coefficient between the scores on the odd and even items of the test. This correlation gives an estimate of the reliability of each half of the test. To obtain the reliability of the whole test, application is made of the Spearman-Brown formula.

Recently, other approaches to the problem of obtaining reliability have been made. A number of methods, both the traditional and the more recent, will be discussed and illustrated in the following pages.

Problem VI.9. Comparison of the split-test and the maximum likelihood methods. We shall compare the results from determining the reliability of a test by the split-test method, using the product-moment

TABLE 33

The Scores of a Random Sample of 25 Students on a Biology Test

Individual	Odd, X	Even, Y
1	227	226
2	124	111
3	210	237
4	178	161
5	192	188
6	104	93
7	191	201
8	148	168
9	125	123
10	141	157
11	171	178
12	168	182
13	129	118
14	192	222
15	176	171
16	172	180
17	215	224
18	102	144
19	177	176
20	109	125
21	146	150
22	180	184
23	179	193
24	141	131
25	141	135
Total	4038	4178

correlation coefficient, with the results obtained from applying the maximum likelihood method. The comparison was made on a test in biology from which the scores of a random sample of 25 students are listed in Table 33.

Before applying the split-test method, it was necessary to test the underlying assumptions, namely, that the means on the two halves of the test are equal and that the standard deviations are equal. The t-test for the former ($t_0 = 1.917$) and the F-test for the latter ($F_0 = 1.238$) give probability values $P > .05$. Therefore we may consider the assumptions satisfied and proceed to determine the correlation between the two halves of the test by calculating the product-moment correlation coefficient and by getting the maximum likelihood estimate.

The product-moment correlation coefficient between the scores on the two halves of the test is given by

$$\begin{aligned} r &= \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \\ &= \frac{29{,}812.44}{\sqrt{(28{,}930.24)(35{,}816.64)}} = 0.9262 \text{ (or 0.93)} \end{aligned}$$

The maximum likelihood estimate is given by

$$\begin{aligned} r' &= \frac{2\Sigma XY - \dfrac{(\Sigma X + \Sigma Y)^2}{2N}}{\Sigma X^2 + \Sigma Y^2 - \dfrac{(\Sigma X + \Sigma Y)^2}{2N}} \\ &= \frac{2(704{,}643) - 1{,}350{,}053.12}{681{,}148 + 734{,}044 - 1{,}350{,}053.12} = 0.9093 \text{ (or 0.91)} \end{aligned}$$

Although the difference between the two estimates in this problem can not be said to be large, we accept the maximum likelihood estimate as the optimum estimate.

Problem VI.10. Comparison of the product-moment correlation coefficient and the maximum likelihood estimate for determining the reliability of a test by means of the equivalent forms method. The comparison of these two methods of estimating test reliability was made on the scores of two forms of a reading test made by a random sample of 30 students. The data are given in Table 34.

It was found that the value given for the product-moment correlation coefficient was 0.9164 or 0.92 and the maximum likelihood estimate was 0.9076 or 0.91. Although the difference in this problem is small, we accept the maximum likelihood estimate as the optimum.

It should be observed that previous to the application of the method of estimation, the fundamental assumptions underlying the equivalent forms method of testing reliability have been tested. The assumption here is that the standard deviations on the scores of the two forms of the

TABLE 34

THE SCORES ON TWO FORMS OF A READING TEST OF A RANDOM SAMPLE OF 30 STUDENTS

Individual	Form B, X	Form A, Y
1	46	39
2	47	45
3	46	42
4	27	35
5	59	53
6	74	64
7	30	27
8	50	41
9	56	50
10	41	43
11	24	25
12	27	32
13	37	34
14	59	54
15	36	38
16	42	42
17	41	45
18	49	39
19	29	28
20	57	50
21	27	26
22	49	46
23	34	26
24	14	23
25	44	50
26	48	46
27	61	64
28	70	69
29	58	49
30	50	60
N = 30 Total	1332	1285

test are equal. The F-test ($F_0 = 1.32$) showed that this assumption was satisfied.

A more stringent test of the equivalence of the two forms of the test can be made by applying three sample criteria proposed by Wilks (Ref. 34) for testing the equality of means, equality of variances, and equality of covariances. The L_{mvc} criterion (two variables) is

$$L_{mvc} = \frac{s_{11}s_{22} - s_{12}^2}{[\frac{1}{2}(s_{11} + s_{22}) + \frac{1}{4}(\bar{X}_1 - \bar{X}_2)^2]^2 - [s_{12} - \frac{1}{4}(\bar{X}_1 - \bar{X}_2)^2]^2} \qquad (6.37)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means; s_{11} and s_{22} are the variances; and s_{12} is the covariance, between the two forms.

Although tests of significance may be made by the use of the prepared tables, an exact level of significance is given by

$$P = L_{mvc}{}^{\frac{1}{2}(N-2)}$$

From the scores of the thirty individuals given in Table 34, we make the following calculations:

$$\begin{aligned}
\bar{X}_1 &= 44.4\\
\bar{X}_2 &= 42.83\\
s_{11} &= \tfrac{1}{30}[64{,}938 - (44.4)^2] = 193.24\\
s_{22} &= \tfrac{1}{30}[59{,}429 - (42.8333)^2] = 146.2751\\
s_{12} &= \tfrac{1}{30}[(61{,}676) - (44.4000)(42.8333)] = 154.0682\\
\bar{x} &= \tfrac{1}{2}(\bar{X}_1 + \bar{X}_2)\\
&= \tfrac{1}{2}(87.23)\\
&= 43.62
\end{aligned}$$

Substituting the values required in (6.37), we get

$$L_{mvc} = .8268$$

and

$$P = L_{mvc}{}^{\frac{1}{2}(N-2)} = (.8268)^{14} = .07$$

Therefore, we conclude that the two forms of the test are parallel or equivalent.

Problem VI.11. Determining the sensitivity of a test. Jackson (Ref. 18) applied analysis of variance methods and the methods of testing statistical hypothesis to the problem of determining the reliability of a test.[3] He treated four different problems: the determinations of (1) the existence of a significant practice effect, (2) whether or not the test measures the capacities of the individuals tested, and the estimation of (3) practice effect, if it is found to exist; and (4) the relative importance of the random errors of measurement with respect to the true measurement of the capacity of the individual. Jackson introduced a new statistic, γ, which he called the *sensitivity* of the test, defined as the ratio of the standard deviation of true scores to the standard deviation of the distribution of errors of measurement.

The method of Jackson is applied to the scores of a random sample of 30 students on two forms A and B of a reading test, the same data as were used in Problem VI.10. The original data and calculations are given in Table 35.

It is assumed that each individual's score on the test is the sum of a number of independent components and that the analysis gives a measure of the influence of each. One component is the difference in ability

[3] The student may find it advantageous to follow through this method after he has studied the analysis of variance (see page 226). For the method of testing statistical hypothesis see page 63.

between the individuals tested. Noting the scores of the individual students in columns (2) and (3), it is observed that the students on the average make higher scores on form B than on form A. Form A was given first, so that this difference is called a measure of the "practice" effect. Even when allowance is made for the influence of practice effect,

TABLE 35

SCORES OF FRESHMAN STUDENTS IN THE COLLEGE OF AGRICULTURE ON FORMS A AND B OF A READING TEST

Student No.	Score on		Sum of scores $X + Y$	Difference in scores, $X - Y$
	Form B, X	Form A, Y		
(1)	(2)	(3)	(4)	(5)
1	46	39	85	7
2	47	45	92	2
3	46	42	88	4
4	27	35	62	−*8*
5	59	53	112	6
6	74	64	138	10
7	30	27	57	3
8	50	41	91	9
9	56	50	106	6
10	41	43	84	−*2*
11	24	25	49	−*1*
12	27	32	59	−*5*
13	37	34	71	3
14	59	54	113	5
15	36	38	74	−*2*
16	42	42	84	0
17	41	45	86	−*4*
18	49	39	88	10
19	29	28	57	1
20	57	50	107	7
21	27	26	53	1
22	49	46	95	3
23	34	26	60	8
24	14	23	37	−*9*
25	44	50	94	−*6*
26	48	46	94	2
27	61	64	125	−*3*
28	70	69	139	1
29	58	49	107	9
30	50	60	110	−*10*
Sum	1332	1285	2617	47
Sum of squares	64,938	59,429	247,719	1015

the scores on the two forms differ considerably. It is assumed that these residual differences are attributable to the errors of measurement of the test used. Possibly other factors exist, such as possible fluctuations in the ability of the individual students and differences between the two forms. Since these factors are not isolated, they are included—if they exist—in the measurement of error. The method used to measure the effect of each of the components is that of the analysis of variance, which consists in breaking up the sum of squares of the deviations about the grand mean into parts assigned to the respective factors. In this way, the importance of the influence of the respective components can be established and conclusions can be made with respect to the value of the test as a measuring instrument.

The calculations involved in the analysis of variance are as follows:

(1) Calculate for each student the sum of the scores and the difference between his scores on the forms as indicated in columns (4) and (5), Table 35.

(2) Calculate the sum and sum of squares of the numerical values in each of the columns (2), (3), (4), and (5), and record these in the two bottom rows of the table. Note the following checks on the calculations:

(a) $1332 + 1285 \equiv 2617$
(b) $1332 - 1285 \equiv 47$
(c) $247{,}719 + 1015 \equiv 2(64{,}938 + 59{,}429)$

(3) Calculate the sum of squares for each component as follows:

(a) For error: $\frac{1}{2}\left[1015 - \frac{(47)^2}{30}\right] = 470.683$

(b) For between individuals: $\frac{1}{2}\left[247{,}719 - \frac{(2617)^2}{30}\right] = 9714.683$

(c) For practice effect: $\frac{1}{2}\left[\frac{(47)^2}{30}\right] = 36.817$

(d) For total: $64{,}938 + 59{,}429 - \frac{(2617)^2}{60} = 10{,}222.183$

These values are then recorded in an analysis of variance table (Table 36).

TABLE 36
ANALYSIS OF VARIANCE OF SCORES OF FRESHMEN ON TWO FORMS OF A READING TEST

Source of variation	Degrees of freedom	Sum of squares	Mean square
Practice effect	1	36.817	36.817
Between individuals	29	9,714.683	334.989
Error	29	470.683	16.230
Total	59	10,222.183	

The following applications can now be made of the results shown in Table 36:

An estimate of the standard error of measurement of an individual score, s_E, is obtained by taking the square root of the error mean square. We get

$$s_E = \sqrt{16.230} = 4.03 \text{ score units}$$

This gives a direct estimate of the absolute accuracy of the measurements.

The next problem is to test whether there is a significant practice effect, that is, if it is significantly different from zero. This hypothesis is tested by calculating first the ratio of mean square due to practice effect to the mean square due to error:

$$F = \frac{36.817}{16.230} = 2.27$$

We then refer to Snedecor's table (Table IV, Appendix) of F with degrees of freedom $n_1 = 1$ and $n_2 = 29$. We find that the 5 per cent point of F is 4.18. Since the observed value of F, 2.27, is less than 4.18, we conclude that there is no significant practice effect.

The next step is to find out whether the test measures sufficiently accurately to distinguish among the individual students. This is determined by taking the ratio of the mean square between individuals to the error mean square. Thus:

$$F = \frac{334.989}{16.230} = 20.64$$

Referring to Snedecor's table with $n_1 = n_2 = 29$, we find that for $n_1 = 30$ and $n_2 = 29$, $F_{.05} = 1.85$ and $F_{.01} = 2.41$. Therefore we conclude, since 20.64 is greater than 2.41, that the two mean squares differ significantly and hence that the test measures with sufficient accuracy to distinguish between the individuals tested.

The next problem is to determine the relative accuracy of measurement, that is, the relation between the magnitude of the errors of measurement and the size of the differences among individuals. This is given by Jackson's measure, γ, called the *sensitivity* of the test:

$$\gamma = \frac{\sigma_c}{\sigma}$$

where σ_c is the standard deviation of the distribution of ability in the population sample, and σ is the standard deviation of the distribution of errors of measurement.

The unique estimate of γ is obtained as follows:

(a) Subtract the error mean square from the mean square between individuals.

(b) Divide the difference by twice the error mean square.
(c) Take the square root of the quotient as an estimate of γ.

From the values in Table 36, we get

(a) $334.989 - 16.230 = 318.759$
(b) $\dfrac{318.759}{2(16.23)} = 9.82$
(c) Estimated $\gamma = \sqrt{9.82} = 3.13$

The confidence interval is set up as follows:

(a) Calculate the ratio of the mean square between individuals to the error mean square, denoted by F.
(b) Obtain the $F_{.05}$ and $F_{.01}$ points of the distribution of F from Snedecor's table.
(c) The lower limit of the interval, using $F_{.01}$ for example, denoted by $\underline{\gamma}$, is given by

$$\underline{\gamma} = \sqrt{\frac{F}{2F_{.01}} - \frac{1}{2}} \tag{6.38}$$

(d) The upper limit of the interval, $\bar{\gamma}$, using $F_{.01}$ for example, is obtained from

$$\bar{\gamma} = \sqrt{\frac{FF_{.01} - 1}{2}} \tag{6.39}$$

(e) We may make the statement that

$$\underline{\gamma} \leq \gamma \leq \bar{\gamma}$$

and the probability that the statement is correct is .98.

For our problem, we get

(a) $F = \dfrac{334.989}{16.230} = 20.64$
(b) $F_{.01} = 2.42$
(c) $\underline{\gamma} = \sqrt{\dfrac{20.64}{4.84} - \dfrac{1}{2}} = 1.94$
(d) $\bar{\gamma} = \sqrt{\dfrac{49.95 - 1}{2}} = 4.95$
(e) $1.94 \leq \gamma \leq 4.95$

Jackson gives the following relation between the sensitivity and reliability coefficient in the population:

$$\gamma = \sqrt{\frac{\rho}{1 - \rho}} \tag{6.40}$$

where ρ denotes the population reliability coefficient.

From Jackson's Table XI (Ref. 18), the values of ρ corresponding to $\gamma = 1.94$ and $\bar{\gamma} = 4.95$ are approximately .80 and .96, respectively. The true values of γ and of ρ are, of course, unknown.

Problem VI.12. Determination of the reliability coefficient by means of the analysis of variance. Hoyt (Ref. 17) developed a formula for estimating the reliability of a test also by means of the method of analysis of variance. The data used in the calculation are the number of correct responses to each item and the score on the test for each individual. The total sum of squares is broken down into three components: (1) between individuals, (2) between items, and (3) residual component or error.

By subtracting the sum of the sum of squares among individuals and among items from the total, the residual sum of squares is used to estimate the discrepancy between the obtained and the true variance.

The necessary calculations are developed as follows. First, set up a table for tabulation of the required data as shown in Table 37; where $s = 1, 2, \cdots, k$; $i = 1, 2, \cdots, n$; k denotes the number of items; n is the number of individuals; X_{si} denotes the score of the ith individual on the sth item, which is presumably 1 or 0.

TABLE 37

TABULATION OF DATA NECESSARY FOR DETERMINING RELIABILITY BY HOYT'S METHOD

Individual	Item	Score
	$1 \quad 2 \ldots\ldots\ldots k$	
1 2	$X_{si} \ldots\ldots\ldots X_{si}$ $\ldots\ldots\ldots\ldots\ldots$	ΣX_{s1} ΣX_{s2}
· · ·		
n	$X_{si} \ldots\ldots\ldots\ldots X_{si}$	$\sum_{s} X_{sn}$
Total	$\sum_{i} X_{1i} \sum_{i} X_{2i} \ldots \sum X_{ki}$	$\sum_{s}\sum_{i} X_{si}$

Let us define:

$$\textit{Grand mean, } \bar{X}_{..}, = \frac{\sum_{s}\sum_{i} X_{si}}{N}$$

where $N = kn$.

$$\textit{Mean of columns, } \bar{X}_{s.}, = \frac{\sum_{i} X_{si}}{n}$$

$$\textit{Mean of rows, } \bar{X}_{.i}, = \frac{\sum_{s} X_{si}}{k}$$

The sum of squares between items is

$$\sum_s \sum_i (\bar{X}_{s.} - \bar{X}_{..})^2 = \frac{\sum_s \left(\sum_i X_{si}\right)^2}{n} - \frac{\left(\sum_s \sum_i X_{si}\right)^2}{N} \tag{6.41}$$

The sum of squares between individuals is

$$\sum_s \sum_i (\bar{X}_{.i} - \bar{X}_{..})^2 = \frac{\sum_i \left(\sum_s X_{si}\right)^2}{k} - \frac{\left(\sum_s \sum_i X_{si}\right)^2}{N} \tag{6.42}$$

Since $X_{si} = 1$ or 0,

$$X_{si} = X_{si}^2$$

and the total sum of squares is

$$\sum_s \sum_i (X_{si} - \bar{X}_{..})^2 = \frac{\sum_s \sum_i X_{si}\left(N - \sum_s \sum_i X_{si}\right)}{N} = \frac{n_1 n_2}{N} \tag{6.43}$$

where $n_1 = \sum_s \sum_i X_{si}$, or the number of correct responses of all individuals on all the items, and n_2 is the number of incorrect responses.

We shall apply this method to an examination in college mathematics consisting of 80 items ($k = 80$) for a class of 119 students ($n = 119$). The calculations—only summary values—are

(1) The sum of squares between individuals:

$$\frac{\sum_i \left(\sum_s X_{si}\right)^2}{k} - \frac{\left(\sum_s \sum_i X_{si}\right)^2}{N} =$$

$$\frac{338{,}042}{80} - \frac{(6216)^2}{9520} = 166.8426$$

(2) The sum of squares between items:

$$\frac{\sum_s \left(\sum_i X_{si}\right)^2}{n} - \frac{\left(\sum_s \sum_i X_{si}\right)^2}{N} =$$

$$\frac{528.634}{119} - \frac{(6216)^2}{9520} = 383.6201$$

(3) The total sum of squares:

$$\frac{n_1 n_2}{N} = \frac{(6216)(9520 - 6216)}{9520} = 2157.3176$$

These values are then recorded in Table 38, Analysis of Variance.

TABLE 38
ANALYSIS OF VARIANCE OF THE SCORES ON A TEST IN COLLEGE MATHEMATICS

Source of variation	D.F.	Sum of squares	Mean squares	*F*	Hypothesis tested
Between individuals	118	166.8426	1.4139 (a)	8.20$^{\textcircled{1}}$	Reject
Between items	79	383.6201	4.8560 (b)	28.17$^{\textcircled{2}}$	Reject
Residual.........	9322	1606.8549	0.1124 (c)		
Total........	9519	2157.3176			

$$F_{\textcircled{1}} = \frac{(a)}{(c)} = \frac{1.4139}{0.1724} = 8.20 \sim P < .01$$

$$F_{\textcircled{2}} = \frac{(b)}{(c)} = \frac{4.8560}{0.1724} = 28.17 \sim P < .01$$

The following uses can be made of the results in the table:

(1) To test the hypothesis that there is no difference between the means of individuals. We calculate the ratio of the mean square due to individuals to the mean square of residual: $F = \dfrac{1.4139}{0.1724} = 8.20$. We then refer to Snedecor's table of F (Table IV, Appendix) with degrees of freedom $n_1 = 118$ and $n_2 = 9322$. We find that the 1 per cent point of F for $n_1 = 100$ and $n_2 = \infty$ is 1.36. We could interpolate to get the value for $n_1 = 118$ and $n_2 = 9322$, but this operation is unnecessary, since it is obvious that the obtained value of F will be much greater than the table value. Therefore, we conclude that the test measures sufficiently accurately to differentiate among individuals.

(2) To estimate the precision with which the test measures, we may compute the reliability coefficient, r_{tt}, as follows:

$$r_{tt} = \frac{a - c}{a} = \frac{1.4139 - 0.1724}{1.4139} = 0.88 \qquad (6.44)$$

A measure of the absolute accuracy of the test is given by the standard error of measurement of an individual score, s_E, where

$$s_E = \sqrt{\frac{\text{residual s.s.}}{\text{D.F. between individuals}}}$$

$$= \sqrt{\frac{1606.8549}{118}} = 3.68 \text{ score units}$$

Problem VI.13. Determination of the reliability of the test by the method of rational equivalence. Kuder and Richardson developed a method of determining the reliability of a test, which they called the method of rational equivalence (Ref. 20). The term "rational equivalence" arises from the conception of a given test as being equivalent to a hypothetical parallel form where every item on the one form is interchangeable with the corresponding item on the other and thus where each pair of items is equivalent with respect to content and difficulty. Furthermore, it is assumed that all corresponding correlations among the items are equal.

A number of formulas representing varying degrees of rigor are presented. Only the one represented for general use is given here [Ref. 20, Formula (20)]:

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2} \tag{6.45}$$

where r_{tt} is the reliability coefficient; n, the number of items; σ_t^2, the variance of the test items; and $\overline{pq}$, the mean variance of the items.

Jackson and Ferguson (Ref. 19) point out that the derivation of Formula (6.45) can be made on the basis of the equivalence assumption only. We present their derivation:

The variance of a test of n items, as a function of the item variances and interitem covariances, is

$$\left.\begin{aligned} s_t^2 &= \sum s_i^2 + 2 \sum_{i,j} r_{ij}s_is_j \\ &\qquad\qquad (i < j) \\ &= n\overline{s_i^2} + n(n-1)\overline{r_{ij}s_is_j} \end{aligned}\right] \tag{6.45a}$$

where s_t^2 = variance of the test
s_i^2 = variance of item i
s_j^2 = variance of the item j
r_{ij} = correlation between items i and j
$\overline{s_i^2}$ = average item variance
$\overline{r_{ij}s_is_j}$ = average item covariance
n = number of items

Assuming the existence of a hypothetical parallel form of the test, also of n items, the variance of the sum of these two tests is

$$s_T^2 = 2n\overline{s_i^2} + 2n(2n-1)\overline{r_{ij}s_is_j} \tag{6.46}$$

where s_T^2 = variance of the sum of scores on the two equivalent forms of the test.

It is known from the correlation of sums that

$$s_T^2 = 2s_t^2(1 + r_{tt}) \tag{6.47}$$

where r_{tt} = correlation between the test and its hypothetical equivalent, or the reliability coefficient.

When the values of s_t^2 in (6.45a) and s_T^2 in (6.46) are substituted in (6.47) and the equation is solved for r_{tt}, we have

$$r_{tt} = \frac{n}{n-1} \cdot \frac{s_t^2 - \Sigma s_i^2}{s_t^2} \qquad (6.48)$$

which formula is identical with (6.45).

It is to be noted that the assumption made in this derivation, that $\overline{r_{ij}s_is_j} = \overline{r_{i'j'}s_{i'}s_{j'}} = \overline{r_{i'j}s_{i'}s_j}$ (that is, that the covariances are on the average equal), is somewhat less rigorous than in the equivalence assumption. In the latter it is specified that $r_{ij} = r_{i'j'} = r_{i'j}$, and $s_i = s_{i'}$, where the primes (′) refer to the hypothetical equivalent form.

As an illustration of this method, we present the results of the administration of an Industrial Relations Classification Test of 100 test items to a college class of 61 students. An analysis of the scores on the test's gave the following values:

Test variance, σ_t^2, = 169.5067

Average variance of the test items, $\overline{pq}$, = .148299

$$\text{Reliability coefficient, } r_{tt}, = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - n\overline{pq}}{\sigma_t^2}$$

$$= \frac{100}{99} \cdot \frac{169.5067 - 100(.148299)}{169}$$

$$= .9\dot{2}$$

Formula (6.45) is not in an efficient form for calculation. Hoyt (Ref. 16) suggests the following variant:

$$r_{tt} = \frac{n}{n-1} \cdot \frac{kSs + S_i - T(T+k)}{kSs - T^2} \qquad (6.49)$$

where T = sum of scores of all individuals

Ss = sum of squares of each of the scores for all individuals

S_i = sum of squares of each of the total correct responses for all items

k = number of individuals taking the test

n = number of items in the test

Applying (6.49) to the data from the above test, we get:

$$r_{tt} = \frac{100}{99} \cdot \frac{62(52{,}734) + 42{,}929 - 1618(1618 + 62)}{62(52{,}734) - (1618)^2}$$

$$= .9\dot{2}$$

Degrees of Freedom

We have used the concept "degrees of freedom" a number of times without defining it. Since it is such a fundamental concept in statistics, we shall try to add to an understanding of it by referring for its interpretation to three analogous settings—physics, geometry, and statistics.

Physical Interpretation. A rigid body which can move about in space without changing the direction of any line in it is said to have a *motion of translation.* It can also turn about any point, say P, without the position of P changing—a motion known as a *motion of rotation* about P. It can again have a motion compounded of a motion of translation and one of rotation.

Take any convenient frame of reference, $O(X_1, X_2, X_3)$ fixed in a rigid body. The position of the rigid body at any instant is defined uniquely by the position of $O(X_1, X_2, X_3)$. We can specify the position of the body axes by six parameters, for example, the Cartesian coordinates α of O, with respect to fixed axes, and the three angular or polar coordinates ϕ of O. Therefore, the rigid body is said to have 6 degrees of freedom. The 6 degrees of freedom correspond to the positional coordinates just specified. Of course, other equivalent sets of coordinates may be taken. However, if a definite relation or relations are fixed or assigned between the six parameters or positional coordinates, then the rigid body is said to be subject to *geometric* or *kinematic constraint* and has less than 6 degrees of freedom. Each restriction reduces the number of degrees of freedom by 1. The fixture of one point of the body would constitute a constraint and reduce the degrees of freedom of the body by 1. Also, a point might be restricted to lie on a curved guide which in turn is constrained to move in a prescribed way. Sliding or rolling contact imposed between the body and either stable or movable guides represents a more general kind of constraint. The constraints may be represented by functional relations connecting the positional coordinates or parameters (Ref. 15).

Geometric Interpretation. The geometric interpretation of degrees of freedom grows out of a consideration of the conceptions derived from the geometry of n-dimensional space. The geometrical or vectorial representation of a sample as a vector[4] with n orthogonal or mutually perpendicular components was introduced into statistics by Fisher (Ref. 8). He carried out the first systematic investigations of the problems underlying the exact sampling distribution of a number of statistics and thus laid the basis for the solution of many theoretical problems of statistical distributions.

It is well known that a one-to-one correspondence may be set up between all real numbers, x, and all points on a straight line. A similar correspondence exists between all pairs of real numbers (x_1, x_2) and all points in a plane; also between all triplets of real numbers (x_1, x_2, x_3) and all points in a space of three dimensions. We may, then, generalize by considering any system of n real numbers $(x_1, x_2, \ldots, x_n)$ as representing a point or vector x in the nth-dimensional (Euclidean) sample space,

[4] A vector is a quantity which has magnitude and direction. It is a matrix consisting of one single row or column.

V_n. A point in a line has freedom of movement in one dimension; that is, it has 1 degree of freedom. A plane has two dimensions and a point on a plane has 2 degrees of freedom. Likewise, in ordinary space of three dimensions, a point in this space has 3 degrees of freedom. Generalizing, a point in n-dimensional space may be said to have n degrees of freedom.

The numbers or values of the respective elements of a sample, x_1, x_2, . . . , x_n, are, then, the coordinates of the sample point P in multiple dimensional space. The dimensionality of the sample point P is the number of observations, n, in the sample. There are n degrees of freedom. However, if a restriction be placed on the sample point, the number of degrees of freedom is decreased by 1; that is, its dimensionality is reduced by 1 and thus becomes $n - 1$. Correspondingly, each additional restriction or section through sample space carries with it an additional reduction in the dimensionality or number of coordinates. Thus to restrict the point in three-dimension space to a surface, one condition is imposed on its coordinates. To restrict a point in space of three dimensions to a curve, it is necessary to subject its coordinates to two independent conditions (Ref. 31).

An illustration of the reduction of dimensionality is given by considering two planes whose equations are

$$2x - y + 3z - 4 = 0 \tag{1}$$
$$2x - y + 5z + 3 = 0 \tag{2}$$

In (1) only two of the values are independent; given $x = 5$ and $y = 12$, the value of z is fixed as 2. Likewise in (2), given any two values for x and y, z is determined. In each case there are 2 degrees of freedom. If restrictions are imposed such that points which lie on both planes are to be determined, then they must lie on the line of intersection of the two planes. These points are determined by solving the equations for x and z in terms of y, or for y and z in terms of x. Thus: $x = \frac{2y + 29}{4}$; $z = \frac{7}{2}$. Now, there is only one independent observation. That is, by selecting values for y and calculating the corresponding values of x and z from these equations, any desired number of points on the line are obtainable. Since there is only one independent variable, the number of degrees of freedom is 1—the point can move up and down the line of intersection. The dimensionality has thus been reduced to 1.

Statistical Interpretation. In its statistical application, the number of degrees of freedom is the number of free variables in the problem or in the distribution of the random variables connected with it. For each restriction imposed upon the original observations, such as in the estimation of a population value from the sample, the number of degrees of freedom is reduced by 1.

It has been previously noted that the unbiased estimate of the population variance from a sample, n, is obtained by dividing the sum of squares of deviations of the individual observations from their mean by $n - 1$, which is the number of degrees of freedom. In this case, it is observed that this is the number of deviations reduced by the number of parameters estimated from the sample and used to establish the point from which the deviations are measured. In this case, the mean is found from the sample, and hence the number of degrees of freedom is one less than the number of observations.

In the case of establishing a regression line among a distribution of observed values, the straight line will fit any two observations with no residuals. Thus, in fitting the least-square line to 25 observations there are 23 degrees of freedom. Two degrees of freedom have been used up in estimating the two parameters in the regression equation (see page 88).

The principle that for each relationship imposed upon the original observations there is a corresponding reduction in the number of degrees of freedom originally available will be found to apply throughout statistical procedures.[5]

PROBLEMS

1. Show that the maximum likelihood estimate of the population reliability coefficient, ρ, for the case of the split-test method is

$$\rho = \frac{2\Sigma X_iY_i - \dfrac{(\Sigma X_i + \Sigma Y_i)^2}{2N}}{\Sigma X_i^2 + \Sigma Y_i^2 - \dfrac{(\Sigma X_i + \Sigma Y_i)^2}{2N}}$$

when X_i and Y_i denote the scores obtained by the ith individual on the odd and even items of the test, respectively; N, the number of pairs of values; and ρ, the correlation coefficient in the sampled population of X and Y.

2. Set up the confidence interval with a confidence coefficient of 95 per cent for ρ, the population correlation coefficient. The sample value is $r = .77$, the correlation between scores on Miller's Analogies Test and the Otis Intelligence Test for a random sampling of 50 graduate students. Any of the following may be used: The exact tables of the r-distribution (David, F. N., *Tables of the Correlation Coefficient*, *Biometrika* Office, London, 1938); the transformation of r suggested by Pillai (Pillai, K. C. S., *Sankhya*, Vol. 7, Part 4, pp. 415–422, July, 1946); or the logarithmic transformation due to R. A. Fisher.

   The probability for the inequality, where a is determined from the normal scale corresponding to a given confidence coefficient α (say 0.95), is

[5] For equivalence of degrees of freedom to orthogonal linear functions, see the discussion of Analysis of Variance, Chapter X.

$$-a \leq [\sqrt{(n-3)}/2] \log \left[\frac{(1+r)(1-\rho)}{(1-r)(1+\rho)}\right] \leq a$$

3. *Given:* $Y_E = .6570X + 33.76$

as the equation for estimating the score on a mid-quarter examination from a knowledge of the score on Miller's Analogies Test.

$$\begin{aligned} \text{Sum of squared } x\text{-deviations} &= 10{,}584.88 \\ \text{Sum of squared } y\text{-deviations} &= 9788.50 \\ n &= 50 \end{aligned}$$

Set up the confidence interval for b_{yx} with a confidence coefficient of 99 per cent. Let β_{YX} be the population value.

$$\text{Variable, } t = \frac{(b-\beta)\sqrt{\Sigma(X-\bar{X})^2}}{s},$$

has Student's distribution with $n - 2$ degrees of freedom.

$$s^2 = \frac{1}{n-2}\Sigma(Y - Y_E)^2$$

4. With the aid of Nair's tables (Ref. 22) find the 95 per cent and 99 per cent confidence intervals from the following values of the median:

(a) Median = 38, $N = 25$ (c) Median = 42, $N = 229$
(b) Median = 18, $N = 25$ (d) Median = 21, $N = 219$

5. Set up the 99 per cent confidence interval for the difference between the percentages given below obtained in two public-opinion polls:

$$n_1 = 3000, \quad p_1 = .52; \quad n_2 = 800, \quad p_2 = .48$$

6. Set up the 98 per cent confidence interval for the difference in percentages obtained on the same sample:

68 per cent answered "yes" $n = 500$
32 per cent answered "no"

7. Plan in advance the size of sample necessary to provide from the sample an estimate of P so that the confidence belt will be of breadth about .05. Take a confidence coefficient of .95. The value of P from the sample is .60. [See also: Finney, D. J., "Errors of Estimation in Inverse Sampling," *Nature*, Vol. 160 (1947), pp. 195–6.]

8. Set up the fiducial limits of the true mean difference based on the data from the controlled experiment given in Problem 2, page 98. Use a fiducial probability of 95.

9. Set up the fiducial limits of the variance of the distribution of differences based on the data in Problem 2, page 98. Use a fiducial probability of 90.

10. On a particular intelligence test a pupil received an I.Q. rating of 98. On this test the standard error of an individual score is 4.51 I.Q. points. Set up the confidence interval for the true score of the pupil, using a confidence coefficient of 95 per cent.

11. Given: $Y_E = .6570X + 33.76$

which is the equation for predicting Y_E, the score on a mid-quarter examination from a knowledge of a score, X, on Miller's Analogies Test.

$$r = .683,\ s_X^2 = 216.018,\ s_Y^2 = 199.765,\ n = 50,\ \bar{X} = 69.32$$

Determine the confidence interval (99 per cent) for mid-quarter score for the following scores on Miller Analogies:

(a) 99; (b) 69; (c) 27.

(d) Explain your answer to (a) above.

12. The following table (based on the 1940 Census) gives the percentage of adults over twenty-five years of age by states who had not completed more than four years of school:

State	Percentage	State	Percentage	State	Percentage
Iowa	4.1	Dist. of Columbia	8.2	Rhode Island	13.7
Oregon	5.2	Ohio	8.4	Maryland	15.3
Idaho	5.2	Nevada	8.8	West Virginia	16.5
Utah	5.5	Colorado	9.0	Florida	18.5
Washington	5.9	Wisconsin	9.4	Texas	18.8
Nebraska	6.0	Illinois	9.6	Arizona	19.4
Kansas	6.1	Massachusetts	10.1	Kentucky	20.2
Vermont	6.1	Michigan	10.2	Tennessee	21.7
Wyoming	7.1	Missouri	10.3	Arkansas	23.1
South Dakota	7.2	North Dakota	10.8	Virginia	23.2
Montana	7.4	Connecticut	11.2	North Carolina	26.2
Maine	7.4	New Jersey	12.0	New Mexico	27.3
Minnesota	7.5	New York	12.1	Alabama	28.9
Indiana	7.7	Pennsylvania	12.3	Georgia	30.1
California	8.1	Delaware	12.9	Mississippi	30.2
New Hampshire	8.1	Oklahoma	13.5	South Carolina	34.7
				Louisiana	35.7

Problem: Set up the tolerance limits for years of schooling of adults (take α = 90 per cent). How may the results be used in analyzing a state's educational program?

13. Students of fiscal policies are invited to study the characteristics and use of grant-in-aid apportionment formulas in relation to setting up tolerance limits. (Cornell, Francis G., "Grant-in-aid Apportionment Formulas," *Journal of American Statistical Association,* Vol. 42 (1947), pp. 92–104.)

14. The following tabular data are to be used for the problems below:

SCORES OF 25 FRESHMAN STUDENTS ON TEST FORMS A AND B OF A SCIENCE READING TEST

Student No.	Score on Form A	Score on Form B
1	18	21
2	33	37
3	38	44
4	29	30
5	64	63
6	74	68
7	33	36
8	72	66
9	58	51
10	56	57
11	28	39
12	71	76
13	53	53
14	39	40
15	37	42
16	29	27
17	58	68
18	20	26
19	65	71
20	28	31
21	16	23
22	50	44
23	29	32
24	46	54
25	36	35

Problems:

(a) Test the equivalence of the forms A and B of the reading test by
  (1) Testing the equality of the standard deviations of the scores on the two forms.
  (2) Testing the equality of means, variances, and covariances of the scores on the two forms.

(b) Determine the reliability of the reading test by calculating the product-moment correlation coefficient.

(c) Determine the reliability of the reading test by getting the maximum likelihood estimate.

(d)* Determine the sensitivity of the reading test.

(e) Calculate the standard error of measurement of an individual score.

* This may be postponed until the analysis of variance method has been studied.

15. The following data are to be used in the problems below:

Scores of a Random Sample of 25 Students on a Comprehensive Examination in College Biology

Student	Score on items	
No.	Odd	Even
1	143	145
2	175	179
3	158	157
4	178	172
5	113	94
6	143	140
7	136	139
8	234	243
9	201	207
10	203	213
11	222	248
12	200	184
13	195	191
14	126	136
15	186	208
16	163	186
17	160	158
18	188	197
19	196	206
20	253	249
21	206	196
22	167	154
23	148	142
24	188	186
25	204	221

Problems:

(a) Before attempting the methods of determining the reliability of the biology test, test the assumption regarding the means and standard deviations on the two halves of the test.

(b) If the assumptions in (a) are fulfilled, determine the reliability of each half of the test by calculating the product-moment correlation coefficient.

(1) What are the assumptions underlying the use of the Spearman-Brown formula?

(2) If the assumptions in (1) are fulfilled, estimate the reliability coefficient of the whole test.

(c) If the assumptions in (a) are fulfilled, calculate the reliability coefficient of the test by getting the maximum likelihood estimate.

(d) Calculate the standard error of measurement of an individual score.

16. Calculate the reliability coefficient for the English examination by using the method of rational equivalence. The examination of 297 items was administered to a group of 209 college students. The following values were computed from the examination results:

$$\bar{X} = 144.58 \qquad s_t^2 = 775.0656$$
$$s_t = 27.84 \qquad \overline{pq} = .245$$
$$n = 297$$

17. Calculate the reliability coefficient for a mathematics test of 75 items administered to 35 students by using the analysis of variance method (Hoyt). The basic data are given in Tables A and B.

TABLE A
NUMBER OF CORRECT RESPONSES TO EACH OF THE 75 TEST ITEMS

Item	f	Item	f	Item	f	Item	f	Item	f	Item	f	Item	f	Item	f
1	22	11	11	21	18	31	7	41	24	51	9	61	16	71	20
2	25	12	25	22	22	32	9	42	20	52	8	62	16	72	16
3	25	13	11	23	23	33	24	43	26	53	18	63	1	73	23
4	24	14	9	24	17	34	19	44	9	54	14	64	8	74	14
5	8	15	17	25	17	35	26	45	16	55	10	65	20	75	2
6	22	16	27	26	9	36	17	46	15	56	6	66	19		
7	27	17	13	27	14	37	15	47	4	57	9	67	11		
8	11	18	19	28	14	38	13	48	31	58	11	68	9		
9	23	19	23	29	16	39	22	49	24	59	7	69	14		
10	16	20	25	30	25	40	18	50	22	60	26	70	19		

TABLE B
TOTAL SCORES OF THE 35 STUDENTS

Score	f	Score	f	Score	f	Score	f	Score	f
55	1	45	1	36	1	30	1	25	2
54	1	44	2	35	3	29	1	24	1
52	2	43	1	34	1	28	2	23	1
50	1	42	1	33	1	27	1	17	1
48	1	41	1	31	1	26	3	16	1
47	1	39	1						

18. (a) Look up in some reference text or texts (Kelley, Truman L., *Fundamentals of Statistics,* for instance) the following methods of estimating correlation:

(1) Biserial *r*
(2) Point-biserial *r*
(3) Biserial phi-coefficient
(4) Correlation for a fourfold point-surface, or the phi-coefficient
(5) Tetrachoric *r*

(6) Coefficient of mean square contingency

(7) Correlation ratio

(b) Specify the types of problems for which each method in (a) is designed.

(c) What assumptions underlie the use of each method?

(1) How may these assumptions be tested?

(d) Which of the approximate measures of relationship are convertible to the product-moment scale, and under what conditions?

19. Evaluate the several statistics that are in use as indices of internal consistency in item analysis.

20. Plan in advance, from the data in Problem 9, Chapter 5, page 100, the size of sample such that the probability will be .95 that the .99 confidence interval of the mean will have a length less than four-score units.

References

1. Alexander, Howard W., "The Estimation of Reliability When Several Trials Are Available," *Psychometrika*, Vol. 12 (1947), pp. 79–99.
2. Cramér, Harald, *Mathematical Methods of Statistics*, Princeton, N. J.: Princeton University Press, 1946.
3. Fisher, R. A., "The Concepts of Inverse Probability and Fiducial Probability Referring to Unknown Parameters," *Proceedings of the Royal Society* (*London*), Series A, Vol. CXXXIX (1933), pp. 343–348.
4. ———, "The Conditions Under Which χ^2 Measures the Discrepancy Between Observation and Hypothesis," *Journal of the Royal Statistical Society*, Vol. 87 (1924), p. 442.
5. ———, *The Design of Experiment.* London: Oliver & Boyd, Ltd., 1935, pp. 195–198, 249.
6. ———, *idem*, pp. 198–201.
7. ———, "The Fiducial Argument in Statistical Inference," *Annals of Eugenics*, Vol. VI (1935), pp. 391–398.
8. ———, "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Independently Large Population," *Biometrika*, Vol. X (1915), pp. 507–521.
9. ———, "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of *P*," *Journal of the Royal Statistical Society*, Vol. 85 (1922), p. 87.
10. ———, "Inverse Probability," *Proceedings of the Cambridge Philosophical Society*, Vol. XXVI (1930), pp. 528–535.
11. ———, "The Logic of Inductive Inference," *Journal of the Royal Statistical Society*, Vol. 98 (1935), p. 39.
12. ———, "The Logical Inversion of the Notion of the Random Variable," *Sankhya*, Vol. 7, Part 2 (1945).
13. ———, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society* (*London*), Series A, Vol. CCXXII (1921), pp. 309–368.
14. ———, "Theory of Statistical Estimation," *Proceedings of the Cambridge Philosophical Society*, XXII (1925), pp. 700–725.
15. Frazer, R. A., Duncan, W. J., and Collar, A. R., *Elementary Matrices and Some Applications to Dynamics and Differential Equations.* London: Cambridge University Press, 1938.

16. Hoyt, C. J., "Note on a Simplified Method of Computing Test Reliability," *Educational and Psychological Measurement*, Vol. I (1941), pp. 93–95.

17. ———, "Test Reliability Estimated by Analysis of Variance," *Psychometrika*, Vol. 6 (1941), pp. 153–160.

18. Jackson, Robert W. B., "Reliability of Mental Tests," *British Journal of Psychology*, Vol. XXIX (1939), pp. 267–287.

19. ———, and Ferguson, George A., *Studies on the Reliability of Tests.* Toronto: Department of Educational Research, University of Toronto, 1941, pp. 107–112.

20. Kuder, G. F., and Richardson, M. W., "The Theory of the Estimation of Test Reliability," *Psychometrika*, Vol. 2 (1937), pp. 151–160.

21. Markoff, A. A., *Calculus of Probability.* Russian editions: II, 1908; IV, 1924; German edition, 1912.

22. Nair, K. R., "Table of Confidence Intervals for the Median in Samples," *Sankhya*, Vol. 4, Part 4 (1940), pp. 551–558.

23. Neyman, J., "Fiducial Argument and the Theory of Confidence Intervals," *Biometrika*, Vol. XXXII (1941), p. 128.

24. ———, *Lectures and Conferences on Mathematical Statistics.* Washington: Graduate School of U.S. Department of Agriculture, 1938, pp. 143–160.

25. ———, "Outline of a Theory of Estimation," *Philosophical Transactions of the Royal Society* (*London*), Series A, Vol. CCXXXVI, (1937), pp. 333–380.

26. ———, "On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Vol. 97 (1934), pp. 558–625.

27. Pearson, Karl, "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That Arises from Random Sampling," *Philosophical Magazine*, Vol. 50 (1900), p. 157.

28. ———, "Regression, Heredity, and Panmixia," *Philosophical Transactions of the Royal Society* (*London*), Series A, Vol. CLXXXVII (1895), pp. 253–318.

29. Savur, S. R., "The Use of the Median in Tests of Significance," *Proceedings of the Indian Academy of Science* (Section A), Vol. V (1937), pp. 564–576.

30. Thompson, William R., "On Confidence Ranges for the Median and Other Expectation Distributions for Populations of Unknown Distribution Form," *Annals of Mathematical Statistics*, Vol. 7 (1936), pp. 122–128.

31. Walker, Helen M., "Degrees of Freedom," *Journal of Educational Psychology*, Vol. XXI (1940), pp. 253–269.

32. Wilks, S. S., "Confidence Limits and Critical Differences Between Percentages," *Public Opinion Quarterly*, Vol. 4 (1940), pp. 332–338.

33. ———, "Determination of Sample Sizes for Setting Tolerance Limits," *Annals of Mathematical Statistics*, Vol. XII (1941), pp. 91–96.

34. ———, "Sample Criteria for Testing Equality of Means, Equality of Variances, and Equality of Covariances in a Normal Multi-variate Distribution," *Annals of Mathematical Statistics*, Vol. XVII (1946), pp. 257–281.

CHAPTER VII

NORMAL AND NORMALIZED DISTRIBUTIONS IN STATISTICS

The assumption that measurements are distributed in normal probability curves underlies much of statistical theory. The mathematical conditions for normality have been determined (Ref. 8). The best evidence of the fulfillment of these conditions in any particular case is that which is available in the observations. Sometimes, then, it is significant to show that observations are normally distributed or at least that the available evidence indicates a high probability of such a distribution.

The Test of the Hypothesis of Normality. Standard statistical methods are available for testing the hypothesis of normality. The chi-square test of the goodness of fit of theoretical normal frequencies to observed frequencies is a general test of the normality of a distribution of measurements. The test based upon the criteria of Pearson is first presented. Two criteria provide the basis of estimating the extent of agreement between an observed distribution and the normal distribution with respect to two characteristics, symmetry and kurtosis.

The criterion for symmetry is $\sqrt{\beta_1} = \sqrt{\mu_3^2/\mu_2^3}$. The criterion for kurtosis is $\beta_2 = \mu_4/\mu_2^2$. For the normal curve, $\sqrt{\beta_1} = 0$, and $\beta_2 = 3$. It is observed that these criteria involve a second, third, and fourth moment. They are not affected by the size of the unit of measurement employed and are measures of the shape of the unimodal frequency distribution. The measurement of the form of variation of the distribution is given in terms of symmetry and kurtosis, or the flatness of the mode.

Pearson's Test of Normality. The steps in the process of fitting the normal curve to a series of observations by the method of moments are described in detail below.

1. Calculate the first four moment coefficients.

(a) Moments about the mean and origin of ungrouped data. If X is the variate, measured from the origin; $\bar{X}$ is the arithmetic mean; and N is the size of the sample; then the sth moment coefficient, μ_s, about the mean is

$$\mu_s = \frac{1}{N} \sum (X - \bar{X})^s \tag{7.01}$$

In practice, usually with machine calculation, it is convenient to calculate first the powers of the observed values of X measured from the origin. Then the sth moment coefficient, μ'_s, about the origin is

$$\mu'_s = \frac{\Sigma X^s}{N} \qquad (7.02)$$

Then the first four moment coefficients about the mean can be found from those about the origin from the following equations:

$$\left.\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \mu'_2 - (\mu'_1)^2 \\ \mu_3 &= \mu'_3 - 3\mu'_1\mu'_2 + 2(\mu'_1)^3 \\ \mu_4 &= \mu'_4 - 4\mu'_1\mu'_3 + 6(\mu'_1)^2\mu'_2 - 3(\mu'_1)^4 \end{aligned}\right] \qquad (7.03)$$

These equations may be obtained by expanding the binomial, $(X - \bar{X})^s$, and finding the mean for each term of the expansion, separately.

(b) Moments from grouped data.

When the original observations are first grouped into a frequency distribution, it is assumed that all values in a class interval have the value of its central point. Thus if n_t is the number of observational values in the tth class interval and X_t is its central value, then the sth moment coefficient, say V'_s, is given by

$$V'_s = \frac{1}{N} \sum_t n_t X_t^s \qquad (7.04)$$

The moment coefficients V'_s for group data should then be reduced to the values V_s about the mean by means of equations as follows:

$$\left.\begin{aligned} V_1 &= 0 \\ V_2 &= V'_2 - (V'_1)^2 \\ V_3 &= V'_3 - 3V'_1V'_2 + 2(V'_1)^3 \\ V_4 &= V'_4 - 4V'_1V'_3 + 6(V'_1)^2V'_2 - 3(V'_1)^4 \end{aligned}\right] \qquad (7.05)$$

2. Calculate the adjustments for grouping errors.

The assumption in grouped data is that the observations take the value of the mid-point of the class interval. This assumption can be more nearly fulfilled if corrections for grouping, known as *Sheppard's corrections*, are applied. No corrections are necessary in the first and third moment, since the effects of grouping tend to balance out. They are made in the second and fourth moments when the statistics are a system of areas and the height of the curve tapers off gradually at both tails. These corrections serve then to give a better estimate of the parameter values. The sth moment coefficients, μ_s, with Sheppard's corrections, are

$$\left.\begin{aligned} \mu_1 &= V_1 \\ \mu_2 &= V_2 - \tfrac{1}{12}(h^2) \qquad (h = \text{length of interval}) \\ \mu_3 &= V_3 \\ \mu_4 &= V_4 - \tfrac{1}{2}V_2(h^2) + \tfrac{7}{240}(h^4) \end{aligned}\right] \qquad (7.06)$$

3. Calculate β_1 and β_2.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}; \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} \tag{7.07}$$

If normal:
$$\sqrt{\beta_1} = 0; \qquad \beta_2 = 3 \tag{7.08}$$

4. Test whether the obtained values of $\sqrt{\beta_1}$ and β_2 differ significantly from 0 and 3.

The exact sampling distributions of $\sqrt{\beta_1}$ and β_2 when the population is normal have not been worked out, but E. S. Pearson (Ref. 18) has determined approximate empirical frequency curves from the moments of the sampling distributions. Tables giving values of $\sqrt{\beta_1}$ and β_2 are available by which it can be determined according to size of sample how much deviation may be expected from 0 and 3 due to random sampling errors alone.

If either one or both of the criteria, $\sqrt{\beta_1}$ and β_2, differ significantly from the values for the normal curve, 0 and 3 respectively, the hypothesis

TABLE 39

THE COMPUTATION OF THE FIRST FOUR MOMENTS FOR USE IN DETERMINING PEARSON'S CRITERIA OF NORMALITY

Group interval	f	$x = \frac{X - 144.5}{10}$	fx	fx^2	fx^3	fx^4
(1)	(2)	(3)	(4)	(5)	(6)	(7)
229.5–239.5	3	9	27	243	2,187	19,683
219.5–229.5	14	8	112	896	7,168	57,344
209.5–219.5	31	7	217	1,519	10,633	74,431
199.5–209.5	50	6	300	1,800	10,800	64,800
189.5–199.5	56	5	280	1,400	7,000	35,000
179.5–189.5	78	4	312	1,248	4,992	19,968
169.5–179.5	75	3	225	675	2,025	6,075
159.5–169.5	81	2	162	324	648	1,296
149.5–159.5	81	1	81	81	81	81
139.5–149.5	81	0	0	0	0	0
129.5–139.5	77	− *1*	− *77*	77	− *77*	77
119.5–129.5	53	− *2*	−*106*	212	− *424*	848
109.5–119.5	46	− *3*	−*138*	414	−*1,242*	3,726
99.5–109.5	31	− *4*	−*124*	496	−*1,984*	7,936
89.5– 99.5	22	− *5*	−*110*	550	−*2,750*	13,750
79.5– 89.5	19	− *6*	−*114*	684	−*4,104*	24,624
69.5– 79.5	15	− *7*	−*105*	735	−*5,145*	36,015
59.5– 69.5	0	− *8*	0	0	0	0
49.5– 59.5	4	− *9*	− *36*	324	−*2,916*	26,244
39.5– 49.5	1	−*10*	− *10*	100	−*1,000*	10,000
29.5– 39.5	1	−*11*	− *11*	121	−*1,331*	14,641
Total	$N = 819$	–	885 Σfx	11,899 Σfx^2	24,561 Σfx^3	416,539 Σfx^4

that the sample could be a random sample from a normal population is rejected.

Problem VII.1. Testing the normality of a sample by Pearson's method. The fitting of the normal curve to a set of observations is carried out on a set of achievement-test scores of 819 students on a final examination in a college course in general zoology. The arithmetical labor is substantially reduced over that of following directly the process specified in Equation (7.04) by taking the origin near the center of the distribution and proceeding to work with the class interval as the unit. This is done by calculating the moments about the origin of the computation variable, x. The corrections indicated in Equation (7.06) can then be made, putting $h = 1$. The whole process is followed out as recorded in Table 39.

We shall follow through the calculations in the order in which they have been presented in the preceding theoretical discussion. The mean and standard deviation of the distribution are as follows:

$$\bar{x} = \tfrac{885}{819} = 1.08059$$

$$\bar{X} = 144.5 + 10.8059 = 155.3059$$

$$s_x = \left(\frac{\Sigma f x^2}{N} - \bar{x}^2\right)^{\frac{1}{2}} = \left(\frac{11{,}899}{819} - (1.08059)^2\right)^{\frac{1}{2}} = \frac{2996.0227}{819} = 3.65814$$

$$s_X = 36.5814$$

Step 1. Calculate moments about the origin of the computation variable:

$$V'_1 = 1.080586 \qquad V'_2 = \frac{11{,}899}{819} = 14.52869$$

$$(V'_1)^2 = 1.1676668$$

$$(V'_1)^3 = 1.26176386 \qquad V'_3 = \frac{24{,}561}{819} = 29.98901$$

$$(V'_1)^4 = 1.36344459 \qquad V'_4 = \frac{416{,}539}{819} = 508.5946$$

Step 2. Calculate moments about $\bar{x}$:

$$V_1 = 0$$

$$V_2 = V'_2 - (V'_1)^2 = 14.52869 - 1.16767 = 13.36102$$

$$\begin{aligned} V_3 &= V'_3 - 3V'_1V'_2 + 2(V'_1)^3 \\ &= 29.98901 - 3(1.080586)(14.52869) + 2(1.26176386) \\ &= 29.98901 - 47.098497 + 2.52352772 = -14.585959 \end{aligned}$$

$$\begin{aligned} V_4 &= V'_4 - 4V'_1V'_3 + 6(V'_1)^2V'_2 - 3(V'_1)^4 \\ &= 508.5946 - 4(1.080586)(29.98901) + 6(1.1676668)(14.52869) \\ &\qquad - 3(1.36344459) \\ &= 476.6694 \end{aligned}$$

Step 3. Correct the moments for grouping by Sheppard's corrections (for computation variable x, we have $h = 1$):

$$\mu_1 = V_1 = 0$$
$$\mu_2 = V_2 - \tfrac{1}{12}h^2 = 13.36102 - .08333 = 13.27769$$
$$\mu_3 = V_3 = -14.585959$$
$$\mu_4 = V_4 - \tfrac{1}{2}V_2h^2 + \tfrac{7}{240}h^4 = 476.6694 - 6.68051 + .029167$$
$$= 470.01806$$

Step 4. Calculate β_1 and β_2 or α_1 and α_2:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-14.585959)^2}{(13.27769)^3} = .09088713$$
$$\sqrt{\beta_1} = \alpha_1 = \frac{\mu_3}{\mu_2^{3/2}} = -.3014$$

We refer to the tables of $\sqrt{\beta_1}$ (Ref. 18) and find that this deviation, $-.3014$, or one greater than this from $\sqrt{\beta_1} = 0$ or $\alpha_1 = 0$ for the normal curve, is to be expected less than once in 100 trials by random sampling from a normal distribution or population. Thus, the distribution under consideration deviates significantly from a normal distribution with respect to $\sqrt{\beta_1}$.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{470.01806}{(13.27769)^2} = 2.666$$

We refer to the tables of β_2 (Ref. 18) and find that the observed value of β_2 or one less than this value is to be expected less than 5 times in 100 trials but more than 1 time in 100 trials in random sampling from a normal population. Thus, the present distribution deviates significantly at the 5 per cent level from a normal population with respect to β_2.

Fitting the Normal Curve to a Set of Observations by the Use of Cumulants. In 1928, R. A. Fisher developed a new kind of symmetric function, the k-statistics, which possess the valuable property of giving particularly simple sampling formulas, obtainable directly by combinatorial methods, and removing most of the algebraic labor characteristic of the older methods. The k-statistics, $k_p(p = 1, 2, \cdots)$, are symmetric in the observations, $X_1, \ldots, X_n$, so that the mean value of k_p is the pth cumulant, or $E(k_p) = \kappa_p$.

Fisher's criteria to test for the departure from normality of an observed sample, known as the statistics g_1 and g_2, are calculated from the k-statistics, k_1, k_2, k_3, and k_4, which are in turn derived from the sums of powers, from the second through the fourth, of the deviations from the mean. The quantity g_1 is essentially a measure of asymmetry or skewness. The parameter γ of which g_1 is an estimate is related to $\pm \sqrt{\beta_1}$ of Pearson's notation as follows:

$$\pm \sqrt{\beta_1} = \gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}} \tag{7.09}$$

The quantity g_2 is a measure of the peakedness or flatness of the curve, that is, its kurtosis. The parameter γ_2 of which g_2 is an estimate is related to Pearson's β_2 in the following way:

$$\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\kappa_4}{\kappa_2^2} \tag{7.10}$$

A convenient way of calculating the k-statistics is to get first a series of values V_1, V_2, V_3, V_4, defined as follows:

$$\left.\begin{aligned} V_1 &= \frac{\Sigma X}{N} \\ V_2 &= \frac{\Sigma X^2}{N} - \bar{X}^2 \\ V_3 &= \frac{\Sigma X^3}{N} - 3\bar{X}V_2 - \bar{X}^3 \\ V_4 &= \frac{\Sigma X^4}{N} - 4\bar{X}V_3 - 6\bar{X}^2V_2 - \bar{X}^4 \end{aligned}\right] \tag{7.11}$$

The k-statistics are then given by

$$\left.\begin{aligned} k_1 &= V_1 \\ k_2 &= \frac{NV_2}{N-1} \\ k_3 &= \frac{N^2V_3}{(N-1)(N-2)} \\ k_4 &= \frac{N^2(N+1)}{(N-1)(N-2)(N-3)}V_4 - \frac{3N^2}{(N-2)(N-3)}V_2^2 \end{aligned}\right] \tag{7.12}$$

If the sums of powers are calculated from group data, Sheppard's corrections for grouping may be applied as follows:

$$k_2' = k_2 - \tfrac{1}{12}; \qquad k_4' = k_4 - \tfrac{1}{120}$$

However, these corrections should be used for purposes of estimation, not for testing significance.

The statistics g_1 and g_2 are given by

$$g_1 = \frac{k_3}{k_2^{3/2}} \tag{7.13}$$

$$g_2 = \frac{k_4}{k_2^2} \tag{7.14}$$

For samples from a normal population, g_1 is distributed normally about 0 with a sampling variance

$$s_{g_1}^2 = \frac{6N(N-1)}{(N-2)(N+1)(N+3)} \tag{7.15}$$

Similarly, g_2 is distributed normally about 0 with a sampling variance

$$s_{g_2}^2 = \frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)} \tag{7.16}$$

Unless the divergence is marked, large samples are required to detect departure from normality, because the exact sampling distributions of the criteria are not known.

Problem VII.2. Testing the normality of a sample by Fisher's method. An example of the method of testing normality by means of the g-criteria is given by applying it to a sample of the honor-point ratios (H.P.R.) of 302 freshmen in the University of Minnesota College of Agriculture. The calculations are set out in Table 40.

We find that $t_{g_1} = \frac{g_1}{\text{S.E.}_{g_1}} = .166$ and $t_{g_2} = \frac{g_2}{\text{S.E.}_{g_2}} = -1.97$. Entering the normal table or the t-table with degrees of freedom $= \infty$, we find that the respective probabilities are .87 and $< .05$. Therefore we may conclude that the hypothesis of normality is rejected at the 5 per cent level.

Special Treatment of Data to Secure Normal Distributions. Two alternatives are open to the research worker if he finds that his data do not conform to a certain model about which considerable is known and by the use of which the analysis is relatively easy to work out. He may develop a new model to which his data may conform, or he may transform his data to make them fit one of the conventional models. The first alternative is often a problem of considerable mathematical difficulty. Hence the second procedure is usually followed.

In particular, the large part of statistical theory is built on the assumptions that the observations are distributed normally and that the variance is constant. It is often important, therefore, for the research worker to show that his measurements are distributed normally or to transform them into a form that is normally distributed, or at least into a form that has the best possible chance of being so distributed. In some cases the normal probability curve gives a very close approximation to the observed facts. Although this is not often the case, it is usually possible to transform the original observations into some function of them so that the function will be distributed normally. In this way the processes in subsequent calculations become simplified and the results more comprehensive in application. For instance, if the mean and standard deviation of the normal distribution are known, the distribution is known exactly. If any obtained distribution of observations is established as normal, then the known properties of the normal model may be applied to it. Tests of significance become more valid and sensitive when the sampling distribution is normalized in case of original skewness.

The linear scale seems to be used in taking observations almost automatically, as if it were the one unique scale used in nature. This scale may often be the most convenient way of representing the original observations, but it need not be for that reason the only way. Should measurements made in one way follow the normal law, other methods

TABLE 40

CALCULATIONS IN TESTING THE NORMALITY OF A DISTRIBUTION BY THE USE OF THE g-STATISTICS

Honor-point ratios (H.P.R.)	f	d	fd	fd^2	fd^3	fd^4
−1.24 to −1.00	7	−7.5*	−52.5	393.75	−2953.125	22,148.4375
−0.99 to − .75	1	−7	− 7	49	− 343	2,401
−0.74 to − .50	5	−6	−30	180	−1080	6,480
−0.49 to − .25	14	−5	−70	350	−1750	8,750
−0.24 to − .00	24	−4	−96	384	−1536	6,144
0.00 to .25	22	−3	−66	198	− 594	1,782
0.26 to .50	30	−2	−60	120	− 240	480
0.51 to .75	31	−1	−31	31	− 31	31
0.76 to 1.00	28	0	0	0	0	0
1.01 to 1.25	28	1	28	28	28	28
1.26 to 1.50	36	2	72	144	288	576
1.51 to 1.75	20	3	60	180	540	1,620
1.76 to 2.00	26	4	104	416	1664	6,656
2.01 to 2.25	7	5	35	175	875	4,375
2.26 to 2.50	14	6	84	504	3024	18,144
2.51 to 2.75	5	7	35	245	1715	12,005
2.76 to 3.00	4	8	32	256	2048	16,384
Total	302		37.5	3653.75	1654.875	108,004.4375

* All cases in interval H.P.R. = −1.00.

$$V_2 = \frac{\Sigma X^2}{N} - \bar{X}^2$$

$$V_3 = \frac{\Sigma X^3}{N} - 3\bar{X}V_2 - \bar{X}^3$$

$$V_4 = \frac{\Sigma X^4}{N} - 4\bar{X}V_3 - 6\bar{X}^2V_2 - \bar{X}^4$$

$$k_2 = \frac{NV_2}{N-1}$$

$$k_3 = \frac{N^2V_3}{(N-1)(N-2)}$$

$$k_4 = \frac{N^2(N+1)}{(N-1)(N-2)(N-3)} V_4 - \frac{3N^2}{(N-2)(N-3)} V_2^2$$

$$g_1 = \frac{k_3}{k_2^{3/2}} j\ g_2 = \frac{k_4}{k_2^2}$$

$$\text{Variance of } g_1 = \frac{6N(N-1)}{(N-2)(N+1)(N+3)}$$

$$\text{Variance of } g_2 = \frac{24N(N-1)^2}{(N-3)(N-2)(N+3)(N+5)}$$

$$\bar{X} = \frac{37.5}{302} = .1241722$$

$$\begin{aligned}
\bar{X}^2 &= .01541873\\
\bar{X}^3 &= .001914578\\
\bar{X}^4 &= .0002377373\\
3\bar{X} &= .3725166\\
4\bar{X} &= .49668878\\
6\bar{X} &= .7450332\\
6\bar{X}^2 &= .09251238\\
\text{Mean} &= .875 + .25(.124172)\\
&= .875 + .031043\\
&= .906
\end{aligned}$$

$$N^2 = 91{,}204$$

$$\begin{aligned}
(N-1)(N-2) &= 90{,}300\\
(\Sigma X)^2 &= 1406.25\\
N^2V_2 &= N\Sigma X^2 - (\Sigma X)^2\\
&= 1{,}103{,}432.50 - 1406.25\\
&= 1{,}102{,}026.25\\
NV_2 &= 3649.09354\\
V_2 &= 12.083091\\
N^2V_3 &= N\Sigma X^3 - 3(\Sigma X)NV_2 - \bar{X}(\Sigma X)^2\\
&= 499{,}772.25000\\
&\quad -410{,}523.01875\\
&\quad -174.61411\\
N^2V_3 &= 89{,}074.61714\\
V_3 &= .976652
\end{aligned}$$

$$\begin{aligned}
N^2V_4 &= N\Sigma X^4 - 4\bar{X}N^2V_3 - 6\bar{X}^2N^2V_2 - \bar{X}^2(\Sigma X)^2\\
&= 32{,}617{,}340.1250 - 44{,}242.3582\\
&\quad -101951.0803 - 21.6826\\
N^2V_4 &= 32{,}471{,}125.0039\\
V_4 &= 356.027422
\end{aligned}$$

$$k_2 = \frac{3649.09354}{301} = 12.12323$$

$$k_3 = \frac{89{,}074.61714}{90{,}300} = .98643$$

$$\begin{aligned}
k_4 &= \frac{32{,}795{,}836.25505}{89{,}999} - \frac{133{,}158.83}{299}\\
&= 364.40223 - 445.34726\\
&= -80.94503
\end{aligned}$$

$$g_1 = \frac{.98643}{42.211} = .02337$$

$$g_2 = \frac{-80.94503}{146.97281} = -.5507$$

$$\text{Variance of } g_1 = \frac{(6.04)(301)}{(303)(305)} = \frac{1818.04}{92{,}415} = .019673$$

S.E. of $g_1 = .1403$

$$\text{Variance of } g_2 = \frac{(24.16)(301)^2}{(299)(305)(307)} = \frac{2{,}188{,}920.16}{27{,}996{,}865} = .078184$$

S.E. of $g_2 = .2796$

$$t_{g_1} = \frac{.02337}{.1403} = .166 \sim P = .87;\ \text{d.f} = \infty$$

$$t_{g_2} = \frac{-.5507}{.2796} = -1.97 \sim P < .05;\ \text{d.f.} = \infty$$

would not be likely to lead to a similar distribution. For example, measurements of the volume of an object might be found to follow a normal distribution whereas measurements of the diameter would not. Here the measurement of the volume would be the more convenient to deal with. Since the method of measurement giving a normal distribution, if it exists, is not known a priori, it is not likely that the appropriate method will be selected to begin with.

The second condition that is often indicated or implied as a necessary condition for the unfettered use of statistics is the stability or at least the predictability of the variance. Methods of measurement or of transformations giving normal distributions are of special significance when the standard deviation is large in comparison with the mean. In cases where the standard deviation is small, the effect of any transformation is less and, when it is very small, negligible. Both a necessary and sufficient condition for the independence of the mean and standard deviation in samples is normality in the parent distribution.

We now consider the nature and use of various transformations designed to normalize or stabilize variates so as to render their distributions more amenable to treatment by statistical methods based on these conditions.

T-Score. In the field of educational psychology, McCall (Ref. 16) converted the raw scores on a mental test of an unselected group of twelve-year-old children to T-scores. This transformation gives a normal distribution of T-scores. The process is illustrated in the transformation of the raw scores of 141 freshmen on a science test (Table 41).

In columns (1) and (2) the raw-score frequency distribution is given. Column (3) gives the cumulative frequency up to the mid-point of the respective raw-score units; for example, in row 1, $N = 133 + \frac{1}{2}(8) = 137$. In column (4) the cumulative percentages are listed; for example, in row 1, $137/N = 137/141 = 97.13$.

The values recorded in column (5) were obtained from the table of areas and abscissas of the normal curve (Table I, Appendix). Thus, in row 1 the abscissa value of a point, such that 97.13 per cent of the total area under the normal curve lies below the ordinate erected at that point, is found from the table to be 1.90.

The T-score values in column (6) are obtained by multiplying each abscissa value by 10 and adding 50 to the product. Thus, in row 1, $10(1.90) + 50 = 69$.

The T-score unit is defined as one-tenth of the standard deviation. The mean of the distribution of T-scores is 50 and the standard deviation is 10.

It is to be noted that measurements of the mental qualities of individuals may be made so that their distribution will be normal within the limits of sampling error. This result can be obtained for a large unse-

TABLE 41
Transformation of Raw Scores on Johnson Science Application Test of 141 Freshman Students to *T*-Scores

Raw score	*f*	Scores lower + ½ those at given score		Values of abscissa in standard measure	*T*-score
		N	Per Cent		
(1)	(2)	(3)	(4)	(5)	(6)
50	8	137.0	97.13	1.90	69
49	8	129.0	91.46	1.37	64
48	7	121.5	86.14	1.09	61
47	8	114.0	80.82	0.87	59
46	8	106.0	75.15	0.68	57
45	7	98.5	69.83	0.52	55
44	6	92.0	65.23	0.39	54
43	5	86.5	61.32	0.29	53
42	7	80.5	57.07	0.18	52
41	6	74.0	52.47	0.06	51
40	6	68.0	48.21	*−0.04*	50
39	5	62.5	44.31	*−0.14*	49
38	9	55.5	39.35	*−0.27*	47
37	7	47.5	33.68	*−0.43*	46
36	7	40.5	28.71	*−0.59*	44
35	7	33.5	23.75	*−0.71*	43
34	7	26.5	18.79	*−0.89*	41
33	5	20.5	14.53	*−1.06*	39
32	5	15.5	10.99	*−1.23*	38
31	6	10.0	7.09	*−1.47*	35
30	4	5.0	3.55	*−1.81*	32
29	3	1.5	1.06	*−2.30*	27

lected homogeneous group of individuals usually by constructing a test or examination comprised of some very easy items, some very difficult items, and many items of average or intermediate difficulty. Of course, a test can be constructed to conform within limits to whatever shape of distribution is wanted by varying the difficulty of the test, the time allotment for administering the test, the system of weighting the scoring of items, the choice of the unit of measurement, and so forth. Furthermore, even if the examinations yield results that are normal for a homogeneous population, the same examination administered to a special group will likely give scores that are skewed, often as a consequence of selection or of the inappropriateness of the examination to the group tested. Whether a normal or some other type of distribution results from the measurements used, it is obvious that whatever knowledge is gained about the distribution, it concerns the distribution of the function of the trait used in the measuring process. This conclusion is valid because the measurement is indirect, that is, through the measurement of a functional relationship, the exact nature of which is unknown. Our measurements are only the manifestation of the underlying trait. The statement that the mental traits of man are or are not normally distributed is unproved and unprovable. No amount of experimentation, for instance, could demonstrate that intelligence is normally distributed. Our knowledge of its distribution relates to the way in which the mathematical function we use in measuring intelligence is distributed. The frequency distribution of Binet I.Q's, for example, for a large homogeneous population is generally held to be normally distributed. However, even here the extreme lower end of the distribution of I.Q.'s is not normal, since there is an excess of individuals with low I.Q.'s (see Ref. 19, page 102). Thus he who makes a test proceeds by first assuming that the trait is normally distributed and then by deriving measurements which will conform to this model. When the raw scores for a particular sample are found to be skew, one means of normalizing them is to convert them to T-scores.

Only when a trait is measurable directly can the true nature of the distribution of the trait become known. Certain biometrical measurements made on random samplings from homogeneous populations may be normal. Wechsler (Ref. 21) collected available data for 89 measured traits and abilities of human beings. Certain linear measurements, such as stature, length of extremities, the various diameters of the skull, and certain of their ratios like the cephalic index, were the only distributions which might be regarded as normal, although even among these there was often considerable asymmetry.

The Use of Probits in Testing the Normality of Transformations. The best method of transformation to secure normalization must usually be determined by trial and error. The success of any particular method can be determined by the application of the standard statistical methods previously described. However, a simple graphical method is available which can be used to find out which transformations are successful and in what respects other transformations are not. The method was developed for dealing with toxicological and other dosage-mortality data, particularly by Gaddum (Ref. 14) and Bliss (Refs. 3 and 4). Their method, that of probits, is first presented in its use for testing the normality of transformations.

The *probit* is defined in terms of the normal equivalent deviation (N.E.D.), and is readily determined for any given percentage from the unit normal curve. The N.E.D. of a given percentage is the deviation (from the mean) equivalent to the given percentage of the area of the curve. In order to make all values positive, the probit is the value resulting from adding 5 to the normal equivalent deviation.[1] The probit values corresponding to given percentages can be read directly from Fisher and Yates's Table IX (Ref. 13). The graphical method consists in plotting the appropriate transformations of the observations as abscissas either on probability paper or against the corresponding probits as ordinates. If the individuals or experimental subjects vary in such a way that the measurements or transformed measurements of the experimental factor are normally distributed, the probit should be a linear function of the measurement or of its transformation. It is usually immediately apparent whether or not the plotted points are randomly distributed about a straight line. When they are so distributed, one can with practice draw a straight line among the points to fit satisfactorily for most practical purposes.

It is possible to fit regression lines, and maximum likelihood estimates of the population parameter values of the mean and standard deviation can be obtained when more precise methods are needed. A straight-line probit graph fitted by eye provides the first approximation. Although graphical analysis is probably the most efficient method for selecting a suitable function, sometimes it is necessary to determine by computation whether a given transformation is effective or, alternatively, whether the departures from another mode of plotting deviate significantly from normality. The standard statistical tests for this purpose, the statistics g_1 and g_2, have been discussed previously. The first, g_1, measures the skewness of the presumed normal distribution and determines whether or not the chief trend of the points is truly linear; g_2 indicates whether the secondary trends and twists about the straight line are statistically significant. With a small number of observations, only large departures from a straight line will be statistically significant. This conclusion will have been recognized as obvious during graphic analysis, so that the computation may then be seldom worth doing. However, when a number is sufficient for making grouping advisable (say 50 or more), the calculations leading to the testing of the agreement between observations and hypothesis may lead to results that are not apparent from inspection.

The principal use of the graphical method just described is limited in its application to data to the percentages corresponding to the values of the variable. However, the graphical method is at times useful even when more complete information is available (Ref. 14). For example, if there are N observations of a given variable, one method is to rank them according to size. Then the smallest observation is assigned a percentage of $\frac{100}{2N}$ and to succeeding observations, percentages of $\frac{300}{2N}, \frac{500}{2N}, \ldots, \frac{(2n-1)100}{2N}$. These percentages are then changed to probits and each

[1] Compare with the T-score, page 158.

individual observation is plotted. When the data become sufficient, they are grouped and added cumulatively and the probits are then plotted against the points separating the groups. When the number of cases in a group is very small, it is preferable to plot the individual readings or to assume an even distribution of the observations over the range covered by the group.

Again, if straight lines fit the data, the distributions are normal. The mean and standard deviation can be estimated fairly accurately from the graph. The reciprocal of the slope of the line gives the estimate of the standard deviation. The mean is the value of the abscissa for which the probit value (as ordinate) is 5. The customary technique for calculating a regression line is not appropriate when the experimental results are of the kind just described. The best estimates of the mean and standard deviation are obtained by using the ordinary methods directly on the transformed observations. When the original observations· are grouped, the most convenient method may be to estimate these statistics from the moments of the distribution.

The method of probits also provides a general graphical method of normalizing distributions which may be applied when the scale on which the experimental results are measured is altogether arbitrary. If a smooth curve is drawn through the points of a random sample of observations plotted against probits, the curve may be used to convert succeeding observations to a scale of probits. These transformed values are necessarily normally distributed. The validity of this procedure requires that the shape of the original curve and the variance of the transformed curve must be stable. An illustration of the application of this principle is given by Ferguson (Ref. 10) in his presentation of methods for the estimation of the limen and precision of separate items of a mental test. Finney (Ref. 11) applied the method of probit analysis to get the maximum likelihood estimates of the two parameters from the data of Ferguson.

The Logarithmic Transformation. It has been found that many moderately skew frequency distributions arising from empirical data or fulfilling certain theoretical conditions are reduced to normal curves when the original observations are transformed to $X = \log X$. A logarithmic transformation of a variable may not only make the distribution more nearly normal but will often stabilize the standard deviation, that is, make it more or less independent of the original variable. This stabilizing tendency holds where it is found that the standard deviation of the original variable is roughly proportional to the mean, or where the variance is proportional to the square of the mean. This fact makes the logarithmic transformation a powerful one. It has also been found useful in dealing with new material whose distribution is unknown (Ref. 6).

There is also the theoretical justification which indicates that the log transformation for most scientific observations is probably preferable to employing no transformation at all. The normal law may predict negative observations. The fact that there are men of more than double the average weight implies the existence of other men with negative weight. In case of scores of enlisted men on the Army Alpha Intelligence Examination, the measures $M - 2$ S.D. and $M - 3$ S.D. give the nonexistent scores of -12 and -49. When logarithmic transformations of the observations are used, this difficulty does not occur. Measurements of the size of small bodies of the same shape may be based on the diameter or on the volume. If the distribution of the volumes is normal, that of the diameters will necessarily be skew, and vice versa. Again, the use of logarithms does away with the difficulty. If the logarithms of the diameters are distributed in a normal manner with a standard deviation λ, the logarithms of the volumes will be normally distributed with standard deviation 3λ (Ref. 14).

The logarithmic transformation, then, should make easy the interpretation of experimental results when the variations are large. It frequently has a double advantage in making experimental results more consistent and in preventing excessive weight from being given to an occasionally large aberrant observation. Cochran (Ref. 6) indicated that the logarithmic transformation made no significant difference when the coefficient of variation was less than 12 per cent. Natural logarithms, preferred by the mathematician, and common logarithms to the base 10, ordinarily liked better by the experimenter, give equally good results. Gaddum (Ref. 14) uses the symbol λ to denote the standard deviation of the logarithm to the base 10. It is worth noting that as a logical consequence of the adoption of the method of logarithmic transformation, the mean of the logarithms (or the geometric mean of the observations, instead of the arithmetic mean) would be regarded as the most probable value.

Gaddum (Ref. 14) gives general formulas for obtaining the mean and standard deviation of the transformed distribution when the original observations have been grouped on an arithmetic scale. These are

$$\bar{X} = \log_{10}\left(\frac{\bar{X}}{\left(1 + \frac{\sigma^2}{\bar{X}^2}\right)^{\frac{1}{2}}}\right) \tag{7.17}$$

$$\lambda^2 = 0.4343 \log_{10}\left(1 + \frac{\sigma^2}{\bar{X}^2}\right), \tag{7.18}$$

where $\bar{X}$ and σ are the mean and standard deviation, respectively, of the original distribution. Gaddum points out that these estimates are reasonably efficient only when λ is less than 0.14 (Ref. 14), when an estimate of λ within 3 per cent can be obtained by dividing the coefficient of variation by 231.

Gaddum proposed to call the distribution of x "log-normal" when the distribution of log x is normal. He reports a number of studies which show that the log-normal distributions have been found in many fields of work. It is also indicated that its use could have facilitated interpretation of data in certain studies in which difficulties were encountered. In Wechsler's study (Ref. 21), for instance, the curves obtained for many of the measurements of human traits were just the kind which are improved by using the logarithmic transformation. Gaddum calculated the values of λ for some of Wechsler's data. For example, the estimated λ's for weight—0.045 and 0.055—are about three times the λ's for height —0.015, 0.0164, 0.0172, 0.017.

Muhsam (Ref. 17) proposes the use of a "log-arith" grid for the study of relative dispersions of distributions. The log-arith grid is a system of rectangular coordinates in which the axis of abscissas is divided logarithmically and that of the ordinates arithmetically. Generally, distribution curves showing equal broadness on a log-arith grid have equal relative dispersions. A broader curve indicates higher relative dispersion while a narrower curve shows a lower one. This form of graphic representation is particularly suitable in the case of log-normal distributions.

The Square Root and Inverse Sine Transformations. The present extensive use of the analysis of variance attaches special significance to the usefulness of transformations when there is reason to suspect that the theoretical conditions for the application of this technique are not fulfilled. These theoretical conditions are that the experimental errors to which the experimental data are subject are normally and independently distributed with the same variance. The logarithmic transformation just discussed equalizes the variance when it is proportional to the square of the mean. Therefore, this transformation is powerful for dealing with quantitative measurements, and it is used as a preparatory step to an analysis of variance when dealing with certain types of nonnormal data. The main objective in the use of this transformation is to ensure that the standard deviation, as calculated from a residual sum of squares, shall be applicable to the various "treatment" means, even though the means are different. The lack of normality of the distribution of the residual errors as observed in practice may be of secondary importance. Curtiss (Ref. 9) indicates that the logarithmic transformation may possibly be more successful in stabilizing the variance than in normalizing the data.

A unit frequently used in expressing results of experimental or other observational data is the *percentage*, such as the proportion of the total number of observations which have a specified quality. Research workers have only recently considered the problem of including in the experimental designs for collecting this type of data an objective estimate of the experimental errors to which the data are subject. The analysis of variance, uniquely fitted to serve this purpose, was not originally planned for use with percentages. The problem was one of discovering a transformation for the original observations which would satisfy the condition of normality of experimental errors required in the analysis of variance. The transformation used for this purpose is known as the *inverse sine function*, $\sin^{-1}\sqrt{x}$.

The inverse sine transformation applies to fractions or percentages derived from the ratio of two small integers, when the experimental errors follow the binomial frequency distribution. Before an analysis of variance is performed, each percentage is changed to an angle θ so that $p = \sin^2\theta$. As the fraction p varies from 0 to 1 or the observed percentage, P, from 0 to 100, the angle θ changes from 0 to 90 deg. In large samples, the sampling variation of P tends to be normally distributed with a variance dependent only on the number of observations on which the percentage is determined. The variance on the new scale is $821/n$. *Fisher and Yates*, in Tables XII and XIII of Ref. 13, provide tables for converting percentages and fractions to degrees.

For the sampling distribution of the estimated percentages or proportions to be normal, the population value of p would be .50. For values of the parameters departing widely from .50, as between 0 and .25 and between .75 and 1.00, the sampling distribution would be highly skew. For determining measures of sampling errors of such distributions, it is necessary to make a transformation of the observational values. The inverse sine transformation is the one used here.

Likewise, for comparing the differences between percentages, particularly where they deviate widely, as when one is in the tail and the other near the center of the distribution, the inverse sine transformation will render them more nearly comparable. Thus, the difference between two percentages P_1 and P_2 would become

$$d = 100\left(\sin^{-1}\sqrt{\frac{P_1}{100}} - \sin^{-1}\sqrt{\frac{P_2}{100}}\right)$$

and

$$\sigma_d = \sqrt{\frac{821}{N_1} + \frac{821}{N_2}}$$

where N_1 and N_2 are the sizes of the samples. Then $X = d/\sigma_d$ is referred to the normal scale.

Zubin (Ref. 22) has provided nomographs for the test of significance between two percentages transformed to the inverse sine function scale.

When the observational data consist of small integers whose experimental errors follow the Poisson law, the square-root transformation, $y = \sqrt{x}$, is used. This transformation is equivalent to the angular transformation at each end of the percentage scale, that is, from 0 to 20 per cent and from 80 to 100 per cent. For a Poisson distribution with mean m, the standard error is the square root of m. Hence, if the treatments in an experiment bring about differences in the values of p and m, they have different variances. With small whole numbers, treatment differences must be large before they can be significant. Moreover, the larger the treatment differences are, the greater the inequality in their variances is likely to be.

The Poisson distribution is skew and hence there is a known relation between the standard error and the mean. The theoretical variance of the transformed values, $\sqrt{x}$'s, is $\frac{1}{4}$. The purpose of the transformation is to change the data to a new scale in which the experimental variance is approximately the same for all plots, thus making possible the use of all in estimating the standard error of any treatment comparison.

Normalizing Transformation for Ordinal or Ranked Data. In some types of experimental data, it may be possible or sufficient only to place a series of magnitudes in order of preference without knowledge of their metrical values. For example, in tests of psychological preferences, individuals may be able to express preferences but cannot assign numerical values to whatever forces may be operative in bringing about such preferences. Likewise, in the standardization of food products, an important factor is the determination of consumer preferences, which may be indicated by the ranking of a given set of products in order of choice.

Where the assumptions underlying the order of ranking are fulfilled, namely, the assumptions that the underlying trait may be regarded as continuous and normally distributed, the transformation of ordinal data to a form that is amenable to further analysis (for instance, to the analysis of variance) sometimes may be definitely advantageous. The transformation needed is one which normalizes the data and can be obtained by assigning to each item in a series of given size a score equal to the expected value for an observation of corresponding rank in a normal population with zero mean and unit standard deviation. Tables have been prepared for series of all sizes from 2 to 50 items. Such a table is given by Fisher and Yates's Table XX (Ref. 13). Table XXI in the same source provides the sum of squares for the transformed score of each individual, substantially reducing the labor involved in running the analysis of variance. This type of analysis makes possible tests of differentiation in preference between classes of subjects of different sex, age, or other characteristics.

Bliss (Ref. 5) gives a complete description of the technique for transforming ranks and of its application to a problem of testing consumer preferences. Sandon (Ref. 19) has prepared a nomograph for the scoring of rank data on school examinations.

PROBLEMS

1. Set up a list of statistical tools that depend for their efficiency upon the fulfillment of the conditions of normality of the measurements of the trait or characteristic in the population sampled.

2. What are the effects of nonnormality on the validity of tests of significance—the z or F test, the two-tailed t-test, the one-tailed t-test?
3. Test the hypothesis of normality of the following distribution of scores on the factual information test of the 1947 Minnesota State Board Examination in Biology administered in a representative sampling of 56 Minnesota high schools (Anderson, 1949). Use the method of Pearson.

Score	*Frequency*	*Score*	*Frequency*
25	1	12	173
24	3	11	159
23	24	10	129
22	26	9	109
21	73	8	49
20	90	7	28
19	122	6	18
18	179	5	11
17	206	4	6
16	227	3	1
15	255	2	1
14	218	1	0
13	240	Total	2,348

4. Test the hypothesis of normality of the following distribution of first-quarter honor-point ratios of a random sample of 122 students in the College of Agriculture of the University of Minnesota. Use the criteria of Fisher.

H.P.R.	*Frequency*
2.76 to 3.00	1
2.51 to 2.75	2
2.26 to 2.50	9
2.01 to 2.25	3
1.76 to 2.00	5
1.51 to 1.75	8
1.26 to 1.50	13
1.01 to 1.25	14
0.76 to 1.00	11
0.51 to 0.75	14
0.26 to 0.50	14
0.00 to 0.25	11
−0.24 to *0.00*	9
−0.49 to *−0.25*	3
−0.74 to *−0.50*	2
−0.99 to *−0.75*	1
−1.24 to *−1.00*	2
Total	122

5. Use the graphical method involving the use of probits for testing the normality of the distribution of honor-point ratios in Problem 4.

References

1. Aitken, A. C., *Statistical Mathematics*, 3d ed. London: Oliver & Boyd, 1944.
2. Bartlett, M. S., "The Use of Transformations," *Biometrics*, Vol. 3 (1) (1947), pp. 39–52.

3. Bliss, C. I., "The Calculation of the Dosage-Mortality Curve," *Annals of Applied Biology*, Vol. 22 (1935), pp. 134–167.
4. ———, "The Methods of Probits," *Science*, Vol. 79 (1934), pp. 38–39; 409–410.
5. ———, "A Technique for Testing Consumer Preferences with Special Reference to the Constituents of Ice Cream." *Connecticut Agricultural Experiment Station Bulletin* 251 (1943).
6. Cochran, W. G., "Some Difficulties in the Statistical Analysis of Replicated Experiments," *Empire Journal of Experimental Agriculture*, Vol. VI. No. 22 (1938), pp. 157–175.
7. ———, "The χ^2 Correction for Continuity," *Iowa State College Journal of Science*, Vol. XVI (1942), pp. 421–436.
8. Cramér, Harald, "Random Variables and Probability Distributions." *Cambridge Tracts in Mathematics No.* 36, London: Cambridge University Press, 1937.
9. Curtiss, J. H., "On Transformations Used in the Analysis of Variance," *Annals of Mathematical Statistics*, Vol. XIV (1943), pp. 107–122.
10. Ferguson, G. A., "Item Selection by the Constant Process," *Pyschometrika*, Vol. 7 (1942), pp. 19–29.
11. Finney, D. J., "The Application of Probit Analyses to the Results of Mental Tests," *Psychometrika*, Vol. 9 (1944), pp. 31–39.
12. Fisher, R. A., "Moments and Product-Moments of Sampling Distributions," *Proceedings of the London Mathematical Society*, Vol. 30 (1928), pp. 199–238.
13. ———, and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research.* London: Oliver & Boyd, 1938.
14. Gaddum, J. H., "Lognormal Distributions," *Nature*, Vol. 156, No. 3964 (1945), pp. 463–466.
15. Johnson, Palmer O., and Tsao, Fei, "Testing a Certain Hypothesis Regarding Variances Affected by Means," *Journal of Experimental Education*, Vol. 13 (1945), pp. 145–149.
16. McCall, W. A., *How to Measure in Education.* New York: The Macmillan Company, 1922.
17. Muhsam, H. V., "Representation of Relative Variability on a Semilogarithmic Grid," *Nature*, Vol. 158, No. 4013 (1946), p. 453.
18. Pearson, Karl (ed.), *Tables for Statisticians and Biometricians*, London: *Biometrika*, University College.
19. Sandon, A., "Scores for Ranked Data in School Examination Practice," *Annals of Eugenics*, Vol. XIII, Part 2 (1946), pp. 118–121.
20. Stevens, W. L., "The Logarithmic Transformation," *Nature*, Vol. 158, No. 4018 (1946), p. 622.
21. Wechsler, D., *The Range of Human Capacities.* Baltimore: The Williams & Wilkins Company, 1935.
22. Zubin, Joseph, "Nomographs for Determining the Significance of the Differences Between the Frequencies of Events in Two Contrasted Series or Groups," *Journal of the American Statistical Association*, Vol. 24 (1939), pp. 539–544.

CHAPTER VIII

STATISTICAL ANALYSIS OF DATA UNDER NONNORMAL ASSUMPTIONS

A type of data met with at times, particularly in psychology, consists of rankings which may arise from material not capable of quantitative measurement on a variate scale but arranged in order according to some qualitative characteristic. This might be, for example, the problem of arranging musical compositions in the order of preference by a group of students. Another problem consists in ranking according to two variables: the arrangement of a set of musical compositions in the order of preference by a group of professional musicians and by a group of laymen. The relationship between the two sets of rankings is of interest. Another type of data in this field would be produced by having a judge rate individuals on a five-point scale according to some trait. Transformations of these types of data are sometimes made. For example, the ranked data may be transformed into normally distributed data as described in Chapter VII. In another method the ranked data are distributed into groups, so that the frequencies in the various groups follow the normal scale. Scores on a linear scale are then assigned to the groups. Further statistical treatment usually follows, such as computing the product-moment correlation coefficient, using multivariate analysis or factor analysis.

Before deciding to make such transformations, the critical investigator will examine his data and the conditions under which they were collected, to determine whether the assumptions underlying the transformation can be reasonably accepted. He may find that they cannot be and hence decide that a transformation is not warranted. There are a number of simpler statistical methods available which do not require the assumptions of the more elaborate methods suggested above. They enable a direct attack to be made on the data. Some of these methods will now be pointed out, particularly those whose usefulness has been enhanced by the development of means for testing their significance.

The Method of Rank Correlation. The rank-correlation method as developed by Spearman is well known. It is recommended, however, that the principal use prescribed for it in elementary texts on statistics be abandoned. This use consists in assuming that the Spearman's rho may be used as a substitute for the product-moment correlation coefficient by the aid of tables usually given to obtain the product-moment equivalent. The formula due to K. Pearson $\left(r = 2 \sin \frac{\pi\rho}{6}\right)$ gives the relationship between the product-moment coefficient, r, and the rank correlation, ρ, when the variates are normal. The assumptions underlying the equivalence, that there are no ties in rank and that the intervals between successive ranks are equal, are not likely often to be found in practice. The use of the rank correlation given here is that as a test of significance.

The Rank Correlation as a Test of Significance. Recent contributions to our knowledge of the rank correlation enable us to use it effectively as a test of the existence of correlation, that is, to test the hypothesis that the qualities under consideration are independent, or rather, that the judgments of them are independent (Ref. 4). Under such conditions, the pairs of rankings of n members drawn at random are independent. Thus, for large numbers of samples, every ranking of one quality will occur in equal frequencies with every ranking of another quality. If one ranking is fixed in the order $(1, 2, \ldots, n)$, it may be correlated with the $n!$ possible permutations of these members. Thus, the exact probability that any correlation result could be due to random sampling errors can be calculated.

This method of the calculation of probability values becomes laborious and practically prohibitive when n is of any substantial size. Olds (Ref. 10), however, has provided tables which give probability values to a close approximation. He tabled the probability values based upon the distributions of $\Sigma(d^2)$. The latter is simply related to r', the rank correlation, by the equation

$$r' = 1 - \frac{6\Sigma(d^2)}{n^3 - n} \tag{8.01}$$

The rank correlation is of special value in testing significance when there is no knowledge of the form of the bivariate distribution or in the case where the form of the distribution is, or is believed to be, non-normal. It should be pointed out that scarcely anything is known about the significance of rank correlation in correlated populations.

Problem VIII.1. Testing the significance of rank correlation. An example is presented to illustrate the test of significance of a rank correlation coefficient by means of Olds's Tables.

The ranks R_1 and R_2 were assigned to 12 individuals with respect to two qualities with the results shown in the table at the top of page 171.

We enter Table V (Ref. 10, page 148) with $n = 12$ and $\Sigma d^2 = 94$; we find the probability of not exceeding 94 by chance is between .02 and .01. Therefore, we may conclude that there is a correlation between the two rankings.

Problem VII.2. Combination of the information from two tests of significance. Another use to which the rank difference correlation may be put is the combination of rank and contingency methods suitable for utilizing simultaneously two kinds of information contained in group data. Table 42, concerning first-year students entering one of the

Individual	R_1	R_2	d	d^2
A	1	1	0	0
B	2	7	5	25
C	3	7	4	16
D	4	2	−2	4
E	5	4	−1	1
F	6	9	3	9
G	7	3	−4	16
H	8	7	−1	1
I	9	5	−4	16
J	10	12	2	4
K	11	10	−1	1
L	12	11	−1	1
Totals			0	94

colleges of the University of Minnesota, gives the number of those who offered two credits in high-school mathematics and those who presented more than two credits in mathematics at the various levels of rating on the College Aptitude Test (C.A.T.).

TABLE 42

FRESHMAN STUDENTS CLASSIFIED ACCORDING TO COLLEGE APTITUDE RATING AND THE NUMBER OF ENTRANCE CREDITS IN HIGH-SCHOOL MATHEMATICS

Units of high-school mathematics	College aptitude percentile rating				Total
	1–25	26–50	51–75	76–100	
(a) Two years..................	67	103	176	127	475
(b) More than two years.........	27	25	39	20	111
Proportion $\frac{(a)}{(a) + (b)}$............	.719	.805	.819	.864	.812
Rank on C.A.T.................	1	2	3	4	
Rank of the proportion..........	1	2	3	4	$r' = +1$

Two tests of significance, independent of each other, are applied to the data: the chi-square test, χ^2, and the rank-correlation coefficient, r'.

The chi-square test of the independence of the principles of classification gives the following results: $\chi^2 = 8.118$, $P = .046$. The rank-difference correlation, r', between the two series, C.A.T. as one variable and the proportion $\frac{a}{a+b}$ as the other variable, gives $r' = +1$, $P = .042$.

To test whether the aggregate of these two tests is significant, we have the following data:

P	$-\log_e P$	Degrees of freedom
.046	3.0791	2
.042	3.1701	2
	Total 6.2492	4

$\chi^2 = 2(6.2492) = 12.4984$; $\sim .01 < P < .02$. The probability of the hypothesis of independence of college aptitude rating and the number of units of high-school mathematics taken (two or more than two) is approximately .014, by interpolation. Interpolation in the χ^2-table for 4 d.f.:

P	χ^2	$\log_{10} P$
.02	11.668	$\bar{2}.30103$
.014	12.498	$\bar{2}.14600$
.01	13.277	$\bar{2}.00000$

Problem VII.3. Analysis of variation by the method of ranks. Friedman (Ref. 1) has developed the method of ranks which was designed to study variation by using ranked data instead of the original quantitative values, avoiding the assumption of normality in the original data. The method can also be used where the available data relate to order only or to a qualitative character capable only of being ranked. This

TABLE 43

RANKS OF PERCENTAGES OF COLLEGE ATTENDANCE FOR SPECIFIED LEVELS OF COLLEGE APTITUDE AND OF SOCIOECONOMIC STATUS

College aptitude intervals	Ranks based on percentage of college attendance by socioeconomic status					
	Below 15	15–18	19–22	23–26	27–30	Above 30
100	6	2	4.5	4.5	3	1
90–99	5	4	2	3	6	1
80–89	6	4	2	3	5	1
70–79	6	4	1	5	3	2
60–69	6	4	5	2	3	1
50–59	5.5	5.5	2	4	3	1
40–49	2	5	3	6	4	1
15–39	4	6	2	5	3	1
(a) Sum of ranks.	40.5	34.5	21.5	32.5	30.0	9.0
(b) Mean rank...	5.063	4.313	2.686	4.063	3.750	1.125
(c) Deviation....	1.563	0.813	−0.812	0.563	0.250	−2.375
(d) Deviation squared....	2.442969	0.660969	0.669344	0.316969	0.0625	5.640625

Theoretical mean = 3.5. Sum of deviation squared = 9.783376.

method makes use of the statistic χ_r^2, which is related to Kendall's coefficient of concordance W (see page 174) as follows:

$$\chi_r^2 = m(n-1)W \qquad (8.02)$$

The distribution of χ_r^2 tends to approach the distributed χ^2, as n tends to infinity, with $(n-1)$ degrees of freedom. Some significance levels of χ_r^2 have been provided (Ref. 1).

The example given in Table 43 shows the procedure of the method of ranks. The data are given by Schultz (Ref. 13).

(1) The ranks were obtained by arranging in ascending order the percentages of male high-school graduates for each row (the college aptitude levels).
(2) The next step was to obtain the mean rank for each column given in line (b).
(3) The third step was to obtain the difference between the mean rank for each column and the theoretical mean 3.5, i.e., $\frac{1}{2}(p+1)$, where p is the number of ranks.
(4) The sum of squares of the differences in (3) was obtained.
(5) Then χ_r^2 was found as follows:

$$\chi_r^2 = \frac{12}{np(p+1)} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} r_{ij} \right)^2 - 3n(p+1) \qquad (8.03)$$

where r_{ij} is the rank entered in the ith row and the jth column; n is the number of ranks averaged. Thus:

$$\chi_r^2 = \frac{12(8)}{6(7)}(9.783376) = 22.365$$

(6) The χ^2-table is entered with 5 degrees of freedom.
(7) Since P is less than .01, it was inferred that there was a significant association between socioeconomic status and college attendance where college ability was controlled.

Wallis (Ref. 14) gives a formula for calculating the statistic, η_r, the rank correlation ratio:

$$\eta_r^2 = \frac{\rho(p+1)\chi_r^2/12}{np(p^2-1)/12} = \frac{\chi_r^2}{n(p-1)}$$

from which
$$\eta_r^2 = \frac{22.365}{8(6-1)} = .5591 \qquad (8.04)$$

and
$$\eta_r = .75$$

Finally, the value of .75 is an estimate of the rank correlation ratio between socioeconomic status and percentage of college attendance when college aptitude is controlled.

The Case of Multiple Rankings. The problem arises in practice of how to determine the agreement among a number of rankings and how to obtain an estimate of a true ranking if a significant concordance among sets of rankings exists. This is the case when there are m rankings of n instead of two. For instance, a group of students might be asked to arrange the photographs of a number of persons unknown to them with respect to their judgments as to the unknown persons' intelligence. It is desired to test whether there is a community of judgments between the students. Of course this experiment is not equivalent to determining a relationship based on order of experimental findings. There could be a substantial agreement about an incorrect order which might be different from the one established by the score of a valid and reliable intelligence test.

Problem VIII.3. Computing and testing the significance of the coefficient of concordance. Let the following represent the rankings of three observers of 8 objects, $A_1, \ldots, A_8$:

Observer	Objects							
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8
1	7	4	2	6	5	3	1	8
2	4	2	1	7	6	3	5	8
3	7	2	1	6	4	5	3	8
Sum of ranks	18	8	4	19	15	11	9	24

The sum of the sum of the ranks of the columns must be 108, that is, $\frac{mn(n+1)}{2}$ where m is the number of observers and n the number of objects. If the concordance were perfect the sums would be 3, 6, 9, 12, 15, 18, 21, and 24, though not necessarily in that order. If there is little or no agreement, the sums are approximately equal. The variance of these sums gives a measure of the ranking concordance.

Kendall (Reference 9, page 411) derives a coefficient of concordance, W, as

$$W = \frac{12S}{m^2(n^3 - n)} \tag{8.05}$$

where S is the sum of the squares of deviations from the mean, $m(n+1)/2$. If agreement is perfect, then the sums of the columns are $m, 2m, \ldots,$

nm and the sum or S is $m^2(n^3 - n)/12$. The range in values of W is from 0 to 1.

In the example above,

$$\text{Mean} = \frac{m(n+1)}{2} = \frac{3(8+1)}{2} = 13.5$$

$$\begin{aligned} S &= (18 - 13.5)^2 + (8 - 13.5)^2 + (4 - 13.5)^2 + (19 - 13.5)^2 \\ &\quad + (15 - 13.5)^2 + (11 - 13.5)^2 + (19 - 13.5)^2 + (24 - 13.5)^2 \\ &= 320.00 \\ W &= \frac{12(320)}{9(8^3 - 8)} \\ &= .85 \end{aligned}$$

To test the significance of an observed value of W, it is essential to determine the distribution of W (or, more conveniently, of S) in the population, which is obtained by permuting the n ranks in all possible ways in each of the m rankings. Kendall (Ref. 6) gives the distribution for some low values of n and m and indicates how to approximate for large values through the use of a continuous distribution. The latter can be done by the use of the z-distribution where

$$z = \frac{1}{2} \log_e \frac{(m-1)W}{1-W} \tag{8.06}$$

and

$$v_1 = (n-1) - \frac{2}{m} \tag{8.07}$$

$$v_2 = (m-1)\left[\left(n - 1 - \frac{2}{m}\right)\right] \tag{8.08}$$

In making this test for low values of m and n, it is desirable to apply the usual correction for continuity by reducing S in Equation (8.05) by unity and increasing the divisor by 2.

We shall illustrate by testing the significance of the obtained value, $W_0 = .85$:

$$W_0' = \frac{(320 - 1)}{\frac{4536}{12} + 2} = .84$$

$$z = \frac{1}{2} \log_e \frac{(2)(.84)}{1 - .84} = 1.1759$$

$$v_1 = \tfrac{19}{3}, \qquad v_2 = \tfrac{38}{3}$$

For $v_1 = 6$ and $v_2 = 13$: $z_{.001} = 1.0306$.
Hence, for $z = 1.1759$ $P < .001$.

The estimate of the true ranking of the objects is intuitively given by taking as rank 1 that object whose sum of ranks is the least. In our problem that object is A_3, followed by objects A_2, A_7, A_6, A_5, A_1, A_4, and A_8. This ranking is obtained by rearranging the 8 totals in rank order.

This solution is given a firmer theoretical basis by showing that it is "best" in a least-squares sense. If any two of the S's are equal, this method is indeterminate, and priority would be given to the object which has the lesser sum of squares of ranks. Where two objects have the same set of ranks, the specific ranking of each can be decided by tossing a coin or by selecting the ranks in a way most unfavorable to the hypothesis under test. An alternative solution might be obtained by splitting the ranks, giving each of the doubtful objects the same rank. This method, however, introduces severe theoretical difficulties in making tests of significance.

The Method of Paired Comparisons. In the method of paired comparison, the observer compares each object with every other one. He indicates which object in a pair he prefers. This method was developed in psychology in the late 1890's. Its use, however, was limited to that of a descriptive statistic. Recently, statistical methods have been developed for testing the consistency of an individual's comparisons and also of the agreement between observers or judges. These developments should enhance the value of the method for research purposes, particularly for the situations for which it has a unique value. In ranking, for example, if the quality under consideration is not measurable on a linear scale, the resulting ranking may give not only a faulty presentation of an observer's preference but also of the variation of the quality in the individuals. Thus in judging preferences in musical composition it is not unlikely that an auditor would judge A as preferable to B, B to C, and C to A. "Inconsistent" preferences of this kind could not occur in ranking, since, if A is placed above B, and B above C, then A is automatically placed above C. Cases also arise in which the judgments of untrained individuals are wanted who might be capable of comparing pairs of individuals with respect to some quality but would not likely be able to rank all the members of even a relatively small group. In animal experiments or in experiments with very young children, for example, in determining food choices, rankings would not be possible. But paired comparisons could be used by presenting the food in pairs and noting which food was eaten first.

Coefficient of Consistence in Paired Comparisons. Kendall (Refs. 6 and 8) gives a method of deriving a coefficient of consistence which indicates how consistent a judge or observer is in making preferences. If an individual observer produces a configuration of inconsistent preferences, the reasons may be that (1) he is incompetent to judge, (2) the differences among the objects may be too small to detect, (3) the attention of the observer may wander during the experiment, (4) the quality under comparison may not be representable by a linear variable.

With n objects, each of the possible pairs, $\binom{n}{2}$, is presented to the subject and his preference of one member of the pair is noted. If the object A is preferred to B, it may be indicated as A → B. In general, if an observer makes preferences of the type A → B → C → D → E → F . . . there is no inconsistency, and this case corresponds to ordinary ranking. The criterion of inconsistence is the "circular" triad. If the n objects are considered as the vertices of a regular polygon of n sides and each vertex is joined with every other one, the direction of the choice can be indicated. Thus, if A is preferred to B, the symbol in the diagram is A → B. Any triangle in the figure in which the arrows all point in the same direction is a "circular" triad. Thus, if an observer makes preferences of type A → B → C → A, the triad ABC is said to be inconsistent.

Kendall (Ref. 6) proved that the maximum possible number of circular triads is $\frac{n^3 - n}{24}$ if n is odd and $\frac{n^3 - 4n}{24}$ if n is even; the smallest number is zero. If d is the number of circular triads in an observed configuration of preferences, he defines ζ, the coefficient of consistence, as

$$\left.\begin{aligned} \zeta &= 1 - \frac{24d}{n^3 - n} \qquad (n \text{ odd}) \\ \zeta &= 1 - \frac{24d}{n^3 - 4n} \qquad (n \text{ even}) \end{aligned}\right] \tag{8.09}$$

From these equations, it is observed that ζ is unity when there are no inconsistencies in the configuration. As the coefficient decreases to zero, the inconsistence, as determined by the number of circular triads, increases.

The next problem is to determine the statistical significance of ζ, that is, to answer the question: With what probability can an obtained value of ζ arise by chance if the judge assigns his preferences at random in relation to the quality under examination?

With n objects, the number of possible configurations of preferences is $2^{\binom{n}{2}}$. Kendall discusses the procedure of investigating the distribution of d in this population of $2^{\binom{n}{2}}$ different members, namely, the method of proceeding from the distribution of n to that for $(n + 1)$. He gives tables with the frequencies and probabilities for the distribution of d for n up to and including 7.

Coefficient of Agreement for m Observers. Kendall (Ref. 6) derived a coefficient of agreement in which the judgments of m observers are obtained by the method of paired comparisons. The coefficient u is given by

$$u = \frac{2\sum_{j}}{\binom{m}{2}\binom{n}{2}} - 1 \tag{8.10}$$

where m = the number of observers, n = the number of objects judged, $\sum_j$ = total number of agreements between judges:

$$\sum_j = \sum \binom{\gamma}{2} \tag{8.11}$$

where γ is the number in each cell.

The coefficient of agreement, u, is unity if and only if there is unanimous agreement in the comparisons. Its minimum value is -1 only when $m = 2$. Kendall gives tables which enable one to make an exact test of significance of u for the following values of m and n: $m = 3$, $n = 2$ to 8; $m = 4$, $n = 2$ to 6; $m = 5$, $n = 2$ to 5; $m = 6$, $n = 2$ to 4. He has also demonstrated that the χ^2-approximation provides an adequate test of significance for values of m and n outside the range of the tables. The expression

$$\left[\sum_j - \frac{1}{2}N\binom{m}{2}\frac{m-3}{m-2}\right]\frac{4}{m-2} \tag{8.12}$$

is distributed as χ^2 where $N = \binom{n}{2}$, with

$$v = \frac{Nm(m-1)}{(m-2)^2} \text{ degrees of freedom} \tag{8.13}$$

Problem VIII.4. Calculating the coefficient of agreement. A class of 67 ninth-grade boys were asked to state their preferences with respect

TABLE 44

Preferences of 67 Ninth-Grade Boys in 9 School Subjects*

Subject	1	2	3	4	5	6	7	8	9	Totals
1. Physical Education	..	41	55	56	58	56	58	57	62	443
2. Industrial Arts	26	..	57	55	57	56	54	60	63	428
3. Literature	12	10	..	28	36	38	36	40	60	260
4. Mathematics	11	12	39	..	29	34	40	37	51	253
5. Social Studies	9	10	31	38	..	34	40	40	51	253
6. Science	11	11	29	33	33	..	36	43	53	249
7. Spelling	9	13	31	27	27	31	..	34	48	220
8. Art	10	7	27	30	27	24	33	..	47	205
9. Composition	5	4	7	16	16	14	19	20	..	101
Total										2412

* This table is read by considering the subject at the left of each row as being preferred γ times over the subject at the top of the column which locates any particular square, where γ is the number in that square. For example, Physical Education is preferred by 41 boys over Industrial Arts.

to 9 school subjects. Each boy was asked to place an × in front of the one member of each of the 36 pairs of subjects which interested him more when he studied it. The preferences are shown in Table 44. The problem is to determine the similarity of preferences among the boys. The measure of agreement is the coefficient of agreement as given in Equation (8.10).

The calculations required are as follows: The calculation of $\sum_j$ as given by Equation (8.11) can be shortened when the objects are arranged in order of total number of preferences by using the following relation:

$$\sum_j = \sum (\gamma^2) - m \sum (\gamma) + \binom{m}{2}\binom{n}{2} \qquad (8.14)$$

where the summation is now carried out over the half of the table below the diagonal. The numbers in this half being smaller than those in the other half, the arithmetic is simpler.

$$\begin{aligned}
\Sigma(\gamma^2) &= \overline{26}^2 + \overline{12}^2 \cdots \overline{20}^2 = 17{,}914 \\
\Sigma\gamma &= 26 + 12 \cdots 20 = 712 \\
m\Sigma(\gamma) &= (67)(712) = 47{,}704 \\
\binom{m}{2} &= \frac{67 \times 66}{2} = 2211 \\
\binom{n}{2} &= \frac{9 \times 8}{2} = 36 \\
\binom{m}{2}\binom{n}{2} &= (2211)(36) = 79{,}596 \\
\sum_j &= 17{,}914 - 47{,}704 - 79{,}596 = 49{,}806
\end{aligned}$$

Hence,

$$u = \frac{2 \times 49{,}806}{\binom{67}{2}\binom{9}{2}} - 1 = 0.2515$$

To test the significance of u, we calculate χ^2 according to Equation (8.12). Thus:

$$\left[49{,}806 - \frac{1}{2}\binom{9}{2}\binom{67}{2}\frac{67-3}{67-2}\right]\frac{4}{67-2} = 653.57$$

is distributed as χ^2 with

$$v = \frac{\binom{9}{2}67(66)}{(67-2)^2}$$

or 37.7 degrees of freedom. The large value of v justifies the use of the normal approximation to the χ^2-distribution. Then

$$\sqrt{(2\chi^2)} - \sqrt{2v - 1} = 4.2$$

This is a highly improbable result on the hypothesis of a random assignment of preferences. Therefore, the coefficient 0.2515 is statistically significant. It may be concluded that there is a certain amount of agreement, though not a strong one, among the boys in their preferences for school subjects.

Problem VIII.5. Measuring the consistency of choices by use of paired comparisons. The distribution of circular triads of a random sample of 15 ninth-grade boys and the coefficients of consistence for preference in school subjects calculated from Equation (8.09) were as follows:

Student Number	d	ζ
1	0	1.000
2	0	1.000
3	0	1.000
4	0	1.000
5	0	1.000
6	0	1.000
7	0	1.000
8	0	1.000
9	1	0.967
10	1	0.967
11	1	0.967
12	1	0.967
13	3	0.867
14	3	0.867
15	8	0.733

For 8 of the boys, there were no circular triads. Therefore, the coefficients were 1,000; that is, $\zeta = 1 - \frac{24(0)}{729 - 9} = 1.00$. For the remaining 7, there were 4 coefficients of value 0.967, 2 of 0.867, and 1 of 0.733.

It may be concluded that these students were able to give a consistent set of choices of school subjects by use of paired comparisons. The reader is invited to validate these conclusions by making the appropriate tests of significance.

PROBLEMS

1. Before an examination, a teacher ranked her class of 25 students according to their expected achievements. After the examination, the rank was determined according to total score. What can be said about the teacher's estimation of the abilities of the students?

Student	Teacher's rank	Examination rank
a	1	5
b	2	1
c	3	9.5
d	4	22
e	5	4
f	6	16.5
g	7	11.5
h	8	19
i	9	9.5
j	10	21
k	11	7.5
l	12	24
m	13	14
n	14	7.5
o	15	2
p	16	3
q	17	25
r	18	6
s	19	16.5
t	20	15
u	21	20
v	22	23
w	23	13
x	24	18
y	25	11.5

2. Combine the information from two tests of significance, the chi-square test, and the rank correlation coefficient applied to the data in Problem 11, Chapter V, page 100.

3. The following tabulation represents the rankings of 5 students based on their preferences for four different musical compositions:

Student	Composition			
	A_1	A_2	A_3	A_4
1	1	3	4	2
2	2	1	3	4
3	2	3	4	1
4	2	1	4	3
5	3	1	2	4

(a) Compute the coefficient of concordance and test its significance.

(b) If a significant concordance among the sets of rankings is found, combine the rankings to obtain the estimate of the true ranking.

4. The following data represent the rankings according to interests in high-school subjects of a random sample of 28 boys in the eleventh grade. The rankings were obtained by three different methods: (1) paired comparison, (2) order of merit, and (3) rating.

Method	Ranks of subjects								
	Phys. Ed.	Ind. Arts	Lit.	Math.	Soc. Sci.	Sci.	Spell.	Art	Comp.
Paired comparison	1	2	3	4	5	6	7	8	9
Order of merit	1	2	4.5	4.5	3	6	7	8	9
Rating	2	1	4	6	3	5	8	7	9

(a) Test the significance of the difference in ranks by the three methods.

(b) If a significant association is found, estimate the amount of association among the three methods.

5. The following tabulation shows the preferences of 67 ninth-grade girls in 9 school subjects:

Subject	1	2	3	4	5	6	7	8	9	Totals
1. Literature	. .	33	41	41	45	48	51	56	60	375
2. Home Economics	34	. .	38	38	41	48	50	50	59	358
3. Physical Education	26	29	. .	28	34	40	46	53	58	314
4. Spelling	26	29	39	. .	34	38	45	46	48	305
5. Mathematics	22	26	33	33	. .	39	45	45	41	284
6. Art	19	19	27	29	28	. .	42	43	44	251
7. Social Studies	16	17	21	22	22	25	. .	36	36	195
8. Composition	11	17	14	21	22	24	31	. .	32	172
9. Science	7	8	9	19	26	23	31	35	. .	158
Total										2412

(a) Compute the coefficient of agreement u.

(b) Test the significance of u.

(c) Compare the value of u for girls with the value of u for boys from the same school given in Table 44.

6. Construct, administer, and analyze the results from a test designed to measure the attitude and its intensity of a specified population toward

some pressing educational issue. (Consult Guttman, Louis, and Suchman, Edward A., "Intensity and a Zero Point for Attitude Analysis," *American Sociological Review*, Vol. 12 (1947), pp. 58–67.)

References

1. Friedman, Milton, "The Use of Ranks to Avoid the Assumption of Normality," *Journal of the American Statistical Association*, Vol. 32 (1937), pp. 675–701.
2. Guttman, Louis, "An Approach for Quantifying Paired Comparisons and Rank Order," *Annals of Mathematical Statistics*, Vol. XVII (1946), pp. 144–163.
3. ———, "A Basis for Scaling Qualitative Data," *American Sociological Review*, Vol. IX (1944), pp. 139–150.
4. Hotelling, Harold, and Pabst, Margaret R., "Rank Correlation and Tests of Significance Involving No Assumption of Normality," *Annals of Mathematical Statistics*, Vol. VII (1936), pp. 29–43.
5. Kendall, M. G., "A New Measure of Rank Correlation," *Biometrika*, Vol. XXX (1938), pp. 81–93.
6. ———, *The Advanced Theory of Statistics*, Vol. 1. London: Charles Griffin & Company, Ltd., 1945.
7. ———, "Partial Rank Correlation," *Biometrika*, Vol. XXXII (1942), p. 277.
8. ———, and Smith, B. Babington, "On the Method of Paired Comparisons," *Biometrika*, Vol. XXXI (1939), pp. 324–345.
9. ———, and ———, "The Problem of *m* Rankings," *Annals of Mathematical Statistics*, Vol. X (1939), pp. 275–287.
10. Olds, E. G., "Distributions of Sums of Squares of Rank Differences for Small Numbers of Individuals," *Annals of Mathematical Statistics*, Vol. IX (1938), pp. 133–149.
11. Rosander, A. C., "The Use of Inversions as a Test of Random Order," *Journal of the American Statistical Association*, Vol. 37 (1942), pp. 352–358.
12. Scheffé, Henry, "Statistical Inference in the Non-Parametric Case," *Annals of Mathematical Statistics*, Vol. XIV (1943), pp. 305–332.
13. Schultz, Frank G., "Recent Developments in the Statistical Analysis of Ranked Data Adapted to Educational Research," *Journal of Experimental Education*, Vol. XIII (1945), pp. 149–152.
14. Wallis, W. Allen, "The Correlation Ratio for Ranked Data," *Journal of the American Statistical Association*, Vol. 34 (1939), pp. 533–538.
15. Wilks, S. S., "Order Statistics," *Bulletin of the American Mathematical Society*, Vol. 54 (1947), pp. 6–50.

CHAPTER IX

SAMPLING THEORY AND PRACTICE

We shall now attempt to make available to the reader some of the results from investigations about sampling from the point of view of their use in the construction of clearer, more concise, and better-organized designs of sampling surveys and experiments. It is expected that the reader will become able to extend and deepen his knowledge of sampling principles by further reading of the more technical accounts and to apply his knowledge to the particular scientific problems in which he is interested. Although our chief interest here is in the empirical or observational parts of applied statistical science, the theoretical part previously developed is basic. Here, as elsewhere in science, both the theoretical and empirical parts are essential: the progress of a science is dependent on their reciprocal influence and simultaneous advancement.

The theoretical part of science is, presumably, based on exact ascertainments, and its purpose is to develop the structure, relationships, and results of hypotheses. The appropriateness and applicability of a conceptual model involve the confirmation or refutation by observation of the hypotheses which enter into the model. The hypotheses must be changed if they are not supported by experience and observation. An adequate scientific methodology evolves through comparisons and evaluations of scientific theories, both from the standpoint of their essential parts and their efficiency in practice. The more explicit the theory is, the more amenable it becomes to the detection of errors or deficiencies that it may possess.

Observation is the basic process of empirical science. The empirical side of science obtains, criticizes, and systematizes the observations. It unites the observations with the theoretical propositions and in this process may reject the hypotheses of the theory, if found necessary. It should be remembered, however, that the empirical side of science is also directed by hypotheses. The specification of the conditions under which the observations are to be made and the form in which they are to be collected are governed or guided by theory. Within the reciprocal relationships, it is probably mutually advantageous that the speculative and the observational sides of science should work somewhat independently, each by its own special method.

Although statistical theory has been concerned chiefly with random sampling, considerable resourcefulness, based perhaps chiefly on common sense and intuition, has resulted in the development of new and effective systems of sampling designs. Much study is being given to the development of needed statistical theory basic to estimating the relative efficiency of different systems of sampling. Sampling is an excellent illustration of the link between theory and practice and of how difficulties are discovered and resolved as they arise in the problems met with in experience.

From an early date, governments have engaged in the collection of statistics of population, commerce, production, consumption, prices, wages, income, and, more recently, with problems of social need and human welfare. Hence, statistics was originally political arithmetic to a great extent. The standard method for the collection of these statistics has been complete coverage and enumeration, of which the classical example is the population census. Theoretically, at least for those population characteristics which remain relatively constant, this procedure appears to be the best. But such an undertaking is costly, difficult to plan and conduct, limited to a relatively few items of information, is time-consuming, and is liable to be out of date by the time the results are published. In fact, the government even with its great resources and facilities, can carry on complete censuses only at rather long intervals. The exigencies of the World War II required the collection of many types of data which could only be done by the use of sample surveys. It is also worth noting that other governmental investigations had at various times resorted to sampling. In the 1940 census, for instance, the Bureau of the Census was able to broaden the scope of its inquiries by including a set of supplementary questions which were answered by a sample of 1 person in 20. Special sampling surveys for securing statistical information are now often made by unofficial agencies and by private individuals, usually to provide the lacking official statistics.

In recent years we have witnessed the extension of sampling methods to a great diversity of situations and for a variety of purposes, for example:

(1) To find out the most efficient particular pattern and location of the observations in an experiment in physics.
(2) To sample a growing crop for studies in plant physiology, agricultural meteorology, and others.
(3) To estimate the amount of acreage devoted to a particular crop or to forecast the expected yields from the economically more important crops.
(4) To investigate the nature and extent of economic and social problems, such as unemployment, housing, delinquency, and crime.
(5) To discover the factors influencing consumers' demands.
(6) To measure public opinion on political, economic, and similar problems; to detect the effects of propaganda.

(7) To determine the frequency distribution of the length of sentences or other factors to characterize the styles of various authors.

(8) To investigate local government by examining the local laws in a few selected years over a 200-year period.

(9) To study methods of technical control in the manufacture of technical products.

(10) To ascertain the location and frequency of individuals having special talents, such as persons able to withstand the rigors of dive-bombing, or individuals with certain types of color-blindness that make them valuable as observers who can detect camouflage.

Most of these investigations would probably be impossible from the standpoint of expense, time, and utility of findings if it were necessary· to investigate the whole field of inquiry in any detail. Furthermore, some investigations require destructive tests; hence, there would be no point to the investigation if the destruction of the whole were essential.

Sampling, of course, is an everyday affair. From time immemorial it has played an essential role in carrying out common human activities. Primitive man who sampled food before he gave it to his children relied on the statistical principle of sampling without knowing that he did so or that such principles exist. The modern housewife relies on the quality of the sample before she purchases in quantity.

Probably because of the rapidly increasing use of sampling in experimentation and in survey studies, rapid development is taking place in the theory and design of sampling investigations.

Sampling Designs. The planning of sampling designs is usually involved in two situations: extensive survey studies, descriptive or analytical; and experimental investigations, which are more restrictive. In both situations the sampling problem is that of securing accurate and representative samples. A representative sample is one in which the measurements made on its units are equivalent to those which would be obtained by measuring all the elements of the population, except for the inaccuracy due to the limited size of the sample.

The principal questions which relate to the setting out of an investigation by sample are

(1) What is the best size of the sampling units?

(2) What number of sampling units should be used to secure the desired degree of precision in the estimates to be made?

(3) What system of sampling will secure the optimum allocation of the sampling units among the population or its subdivision?

Population. To answer these questions, certain assumptions about the unknown population must be made. It is fundamental to use methods of sampling and of estimation that are based on a minimum of unavoidable assumptions and that also make unambiguous their exact implications. It may be stated in advance that there is not one faultless method of sampling. The method to be used is contingent upon the nature of the material available and obtainable for the particular problem under investigation.

In practice, most populations are finite in character; the universe is comprised of a finite number of members. The conditions of an infinite universe, one which contains an infinite number of members, is assumed to be fulfilled in practice by sampling with replacement. A large part of statistical theory is also built on the assumption that the universe is continuous, that the members or some measurable variable make up a continuous set.

A population is called *existent* if all members can be enumerated or if the members can be designated by a law of formation. For instance, the inhabitants of the United States and the universe of positive integers are existent universes. In cards and dice games and roulette, potential universes consist of the millions of combinations of 52 cards, of the millions of throws of a six-sided die, and the millions of turns of a roulette wheel with its 37 numbers. These need only be imagined as hypothetical universes. Likewise, a population of experiments is a hypothetical universe.

The usual practice of the statistician is to refer to the bulk that is being sampled as the *population*, the *universe*, or the *supply*. The choice of a population or universe is a necessary first step in an investigation based on samples. The definition of the population to be covered in the investigation is an integral part of the statement of the purpose of the study.

Randomness. The concept of randomness is fundamental in sampling theory and practice, but it is rarely if ever defined, except perhaps in mathematical language of which the following is illustrative: "A sequence of variates $x_1, \ldots, x_n$ is said to be a random series, or to satisfy the condition of randomness, if $x_1, \ldots, x_n$ are independently distributed with the same distribution; i.e., if the joint cumulative distribution function (c.d.f.) of $x_1, \ldots, x_n$ is given by the product $F(x_1) \ldots F(x_n)$ where $F(x)$ may be any c.d.f." (Ref. 10).

Restricted to the written word, the condition of randomness seems to be based on certain intuitive principles which give practical results. Randomness is a fundamental idea in connection with the selection of values of a variate from a population. The principle is implied in the criterion of random sampling that every member of the population should have an equal and independent chance of being included in the sample.

Tests of randomness are of the greatest significance, since statistical inference is strictly valid only for random samples. It is also a matter of great practical and scientific importance to determine whether the fluctuations manifested by a series of observations are random in character or whether they may be assumed to be the outcome of some factor operating under a definite law.

Testing for randomness is an important problem in quality control of manufactured products and also of special importance in the analysis of time series. The need for such tests has resulted in considerable research for criteria of randomness.

Bias. If a sample has been chosen from a population in such a way as not to be a random sample, then no valid estimate can be made from it of a population parameter.[1] If a sample has been selected by a random method, it gives a result that progressively approaches the population value as the sample is increased in size, assuming that an unbiased method of estimation has been used. If the results obtained are too high or too low, then the sample is called *biased.* The difference between the value determined by a very large *sample* and the parameter or population value is termed an *error of bias.*

Errors of bias follow no known laws by which their amount might be estimated. Errors of bias are incorporated, therefore, with random errors and may thus result in spurious estimates of the latter. In sampling designs every caution is necessary to avoid errors of bias. Even if an efficient method of sampling has been used, errors of bias may arise in a number of ways. For instance, biases have been observed in sampling surveys of households where nobody was found at home when the interviewer called for the first time. The smaller the family, the smaller are the odds that some one will be at home. Unless the visits are continued until complete enumeration is obtained, errors of bias will arise in connection with size of families and other characteristics associated with it. Other instances of bias in sample surveys may be traced to factors such as bias and irregularity in the interviewer, imperfections in the design of the questionnaire, and errors arising from nonresponse on the part of the interviewee.

A classical example of bias arising from an unrepresentative selection of respondents and from the erroneous belief that a large sample could overcome such an error is furnished by the attempt of *The Literary Digest* in 1936 to predict the results of the Presidential election. Approximately ten million post cards were mailed to people whose names were listed in telephone directories and in files of owners of automobiles. Of the 2,350,176 replies received, only 40.4 per cent were in favor of Franklin D. Roosevelt for President. In the election, he received 60.7 per cent of the votes cast. The error of bias was, therefore, approximately 20 per

[1] In systematic sampling, for instance in stratified sampling, the number of elements to be selected from any stratum must be selected at random.

cent. The sample was biased in that the respondents did not constitute a random sample of those citizens who voted in this election.

Questionnaire studies in which the sample selects itself, voluntary replies to requests for opinions on some controversial issue, and letters written to editors of newspapers—all are likely to represent mainly persons who have strong views on the issues one way or another.

Systems of Sampling

The origin of the sampling problem is in the necessity of estimating certain characteristics of a population usually so large that it is practically impossible to examine every member of the population, or so large that the time and cost required to do so would prohibit the undertaking. In this undertaking, it is essential to consider how best to take the sample and to obtain the estimates, and with what precision the estimates have been made. The fundamental statistical problem is, therefore, that of estimation.

Unrestricted Random Sampling. A particularly simple form of sampling technique is illustrated by the classical urn problem. By counting the number of balls of each color in the sample drawn from the urn, the relative proportion of balls of different colors in the sample is determined. From these proportions the color composition of the balls in the urn is inferred. By using the properties of the familiar binomial or multinomial distributions, the margin of error of the estimate can also be calculated.

An analogous situation in principle might be the estimation of the occupational classification of the from 16 to 17 millions of men, twenty-one to thirty-six years of age who in 1940 registered in accordance with the Selective Service Act. Let us assume that each individual had a registration number which was written on a paper and enclosed in a separate capsule and that all capsules were placed in a container utilizing compressed air to secure a constant rotation. One thousand capsules would be drawn at random and the corresponding occupations ascertained. In order that statistical principles might be used in a valid way, it is fundamental that each member of the sample should be chosen strictly at random, which means a method of selection by which each member of the population has an equal and independent chance of being included in the sample, and that the method of selection is completely independent of the characteristics to be examined. This is the method of purely random sampling, sometimes called *unrestricted* or the *unitary unrestricted* type of sampling. This method is regarded as being capable of giving the most accurate results in cases where the elements of the statistical population have equal chances of inclusion and where there is no prior knowledge of the population sampled to provide a basis for selecting individuals.

Systematic Sampling Methods. In contrast to the method of simple random sampling, a number of methods have been developed which may be called *systematic methods.* These methods utilize prior knowledge of the individuals comprising a universe with the view to increasing accuracy and representation of samples. They generally use more complex forms of random sampling called *representative sampling.*

Stratification. One of these systematic methods is based on the use of knowledge of population characteristics, first to divide the population into more homogeneous groups or strata and then to select at random the sampling units from each of these groups. This method has been called *restrictive random sampling* or the *method of stratification.* It is in effect a weighted combination of random subsamples. Various principles have been used to distribute the sampling units among the several strata. One, called *stratified proportionate sampling,* is based on the distribution of sampling units purely proportional to the total number of units in each stratum. In simple random sampling this proportion is left to chance. Another basis is to take the number of sampling units per stratum proportional to the product of the number of sampling units in the stratum by their standard deviations.

Stratified sampling is used in the Gallup polls of public opinion in order to secure representative proportions of various classes of people rather than to rely on the chance determination of these proportions. In the interviews that are made, each subject supplies sufficient information about himself to permit classification according to (1) part of the country, (2) the urban or rural district, (3) socioeconomic status, (4) political affiliation, (5) age, (6) sex. The particular type of stratification used depends on the problem under inquiry.

While some progress has been made, the methods in use for predicting elections are not yet scientific. Among other hazards, the sample design may reflect erroneous judgments as to the factors (used for controls in stratification) truly associated with the characteristic under investigation. Serious biases may also be introduced because the selection of the sampling units within a stratum to be interviewed is not done at random, making it impossible to obtain an unbiased measure of sampling error from the internal evidence of the responses themselves. Furthermore, the population composed of eligible citizens who subsequently go to the polls and vote is difficult, if not quite impossible, to specify in advance of sampling and the trait itself is susceptible to change without notice.

Cluster Sampling. The method of stratified sampling is also used where the unit of sampling is a group rather than the individual. This method, sometimes called *cluster sampling,* is especially important in the study of human populations when the individuals are often grouped (as by families, inhabitants of single houses or apartment houses or of blocks, and so on) as in the census, for instance, and it becomes very difficult to sample individuals at random under such circumstances. Most uses of this method apply a system of "exclusive units," where no individual or group is included in more than one sampling unit. Mahalanobis (Ref. 19) has used a variant of the method called the "zonal configurational" type or the "overlapping system of grid sampling," in which the same individual or group may form a part of more than one sampling unit. He points out that this method is analogous to sampling from an urn with replacement.

Purposive Selection. A method of systematic sampling essentially different in principle is that which is called *purposive selection.* Instead of making a random selection of the sampling units within strata, this method selects such groups of units that have the weighted sample means of certain characteristics, the *controls,* in close agreement with the population values. This method might save time and labor at times. However, it has often proved to be very hazardous and inaccurate, probably because the sampling units are large and few in number, so that it is difficult to secure a representative sample. Furthermore, the method hypothesizes a considerable knowledge of the population in advance of the sampling process. This information is not often available, and it has been found in a particular case that the facts about the population needed for controls served only for the particular year when the sampling survey was made (Ref. 23).

Applications of the purposive method have been made in certain economic surveys by selecting so-called "typical" counties. The practice of selecting a particular school or groups of schools in which experiments are conducted may also be illustrations of this method, especially if general conclusions are drawn for the educational factors under investigation.

Double Sampling. A method of systematic sampling designed especially for sampling human populations is the one called *double sampling* (Ref. 21). This method involves two sampling investigations. The first consists in drawing a large unrestricted sample from the population, determining for each individual the value of the character, the collection of information on which is easy and relatively inexpensive. This secondary character is known to be closely correlated with the primary character with which the investigation is concerned. The collection of data concerning the values of the primary character is expensive. The second investigation consists in drawing a small sample in which the values of both the primary and secondary characters are ascertained. In this method, discussed by Neyman (Ref. 21), the large sample is used to stratify the population into groups within which the secondary character is relatively homogeneous. Since the two characters are highly correlated, this procedure will also result in an effective means of stratification with respect to the principal character. It is possible, therefore, to proceed with the drawing of the small sample out of the strata comprising the large sample. Accordingly, a more accurate estimate of the primary character may be expected to be obtained from the stratification based on the first investigation. The first sample must be large enough to provide an accurate estimate of the population numbers if increased accuracy of estimation is to result through the double sampling method.

A variant of this method is to find the regression of the primary on the secondary character from the data in the small sample. The predicted value in the regression equation which corresponds to the mean value of the second factor in the large sample is then used to estimate the mean value of the primary character for the total population (Ref. 1).

Subsampling. Cochran (Ref. 1) describes a method called *subsampling,* in which a sampling unit may itself be enumerated by subsampling. There might be a hierarchy of sampling units in multistage sampling; for example, sampling units might be selected in the first stage of randomization, within each such selected unit. Smaller sampling units then might be selected by another act of randomization, and so forth. This special form of subsampling has been called "nested" sampling by Mahalanobis (Ref. 19).

THE SELECTION OF THE SAMPLING SYSTEM

No simple principle exists which leads the investigator uniquely to the selection of a system of sampling. From the many sampling designs that can be constructed in order to answer the questions which prompted the research, one will be selected for application on the basis of the nature of the problem, the resources and the materials available or obtainable, and certain statistical and administrative considerations.

From a statistical standpoint, the problem is to secure the best estimate of the population characters chosen for study. On the basis of knowledge of limiting distribution theory and of best linear unbiased estimates, it is the usual practice to take the standard deviation of the sample estimate about the character estimated as the measure of sampling error. The relative efficiency of different methods of estimation is obtained from the ratios of the reciprocals of the variances of sample estimates of the mean. The statistical criterion of efficiency is usually not the only basis of deciding upon the sampling plan. Another principal consideration is the cost of the investigation.

The basis of planning, therefore, is the selection of a sample design which combines precision of the results and expenditures in such a manner that either the cost is a minimum for any specified precision or the precision is a maximum for any assigned cost. Considerable work has been done in recent years on the study of costs associated with the various sampling and estimating operations, including the determination of the relative magnitudes of variances and covariances between and within various kinds of sampling units.

Thus, although no complete theory with practical applicability is available whereby the investigator always could be certain of selecting the "best" sampling design and at the same time the "best" process of estimation and allocation of sampling units, considerable empirical and scientific knowledge is available upon which an intelligent selection can be made. To a certain extent each field of study may have its own peculiar sampling problems. But the principles so far educed have wide and general application. Often an exploratory or pilot investigation may save a good deal of time and unnecessary expense by providing useful information of the cost and variance, or error functions. In addition, the exploratory period can be used advantageously in giving training to workers in both field and statistical work and thus in controlling mistakes and errors arising from the human factor.

STATISTICAL ASPECTS OF SAMPLING DESIGNS

The statistical planning of the program for obtaining observations from samples involves the problems of specification and estimation. A knowledge of the mathematical form of the population is known or assumed to be known, but the values of one or more parameters entering into the form are unknown. Estimates of one or more parameters are desired, each with minimum sampling error.

In most statistical investigations by sample, a central problem is to ascertain the value of an average (Ref. 5).

Consider a population π with a parameter of location μ and of dispersion σ. A sample $X_1, X_2, \ldots, X_n$ is drawn. A function of these X's, say μ', where

$$\mu' = \mu(X_1, X_2, \cdots, X_n) \tag{9.01}$$

is said to be a "mathematical expectation estimate" of μ, if the mean value of μ' in repeated samples is equal to μ. Further, the estimate of μ' may be said to be the best linear estimate of μ, if it is linear with respect to the X's:

$$\mu' = c_1X_1 + c_2X_2 + \cdots + c_nX_n + c_0 \tag{9.02}$$

and if its standard error is less than that of any other linear estimate of μ'.

The value of an average is

$$\bar{X} = \frac{\sum\limits_k \sum\limits_i (X_{ki})}{\sum\limits_k (n_k)} \tag{9.03}$$

where n_k is the number of sampling units in the kth stratum; X_{ki}, the value of the variate in the ith element of the kth stratum; $\Sigma(n_k)$ may be known or unknown, finite or infinite. The major sampling processes such as random sampling with or without replacement, stratified random sampling of individual elements, stratified random sampling of groups or clusters, double sampling, and purposive sampling can be illustrated and differentiated by the different grouping methods for each of which the sum of $\sum_k \sum_i (X_{ki})$ in Equation (9.03) is obtainable.

Insight as to the arrangement of strata and the average to compute has grown out of the study of the problem of estimation in stratified sampling of groups. In stratified sampling, (9.03) becomes

$$\bar{X} = \frac{\Sigma(n_k)(\bar{X}_k)}{\Sigma(n_k)} \quad . \tag{9.04}$$

where $\bar{X}_k$ equals the average value of X in the kth stratum. In some problems, it has been found, that, by choosing the strata so that the regression of $\bar{X}_k$ on some appropriately selected variate Y is linear, an improved estimate of $\bar{X}$ can be made (see Double Sampling, above).

In general, there is no unique unbiased estimate of a parameter. Under particular conditions the best estimate can be found if the quantity is a linear function of the observations as in (9.02) above. A method and the conditions are given in a theorem by Markoff (Refs. 5 and 22).

It is possible to make the obtained estimate the best linear estimate if another stipulation about the variation of the $\bar{X}_k$'s in strata corresponding to different fixed values of Y and n_k is fulfilled. Neyman (Ref. 22), basing his method on Markoff's theorem, has indicated that the numbers in the sample should be proportional to the product of the number of sampling units in a stratum by the standard deviation of the measured character within the stratum. The "best" estimate is defined by the two conditions that (1) it should be a linear unbiased estimate with (2) minimum variance (see Equation 9.02).

A fundamental condition in the best solution is that the total number of sampling units must be kept constant. In Neyman's method the best solution depends on a knowledge of the population standard deviation of each stratum. Sukhatme (Ref. 26) investigated the effect of estimating the standard deviation of the different strata by a preliminary inquiry. He concluded that a gain in efficiency takes place even in the case where the population standard deviations, σ_i's, are estimated from the sample standard deviations, the s_i's, that is, when the σ_i's are different in different strata.

Mahalonabis (Ref. 19) discusses in detail the statistical planning involved in large-scale sample surveys taking into account both the cost and variance functions for obtaining optimum solutions. A comprehensive and critical review of recent statistical developments in sampling and sampling surveys has been made by Yates (Ref. 28).

TYPES OF ERROR IN INVESTIGATION BY SAMPLE

Statistical data are the raw material of judgments, comparisons, and truth. The highly condensed form to which the original data are usually reduced by processes of statistical reduction gives to the final results a display of exactness that is not necessarily intrinsic. In viewing the final product, one should not forget the original material from which it came. In order to evaluate the findings from an investigation, much information is necessary as to the ways in which the original data were collected, the conditions surrounding them, and the kinds of errors to which they are susceptible. We wish to consider here the types of errors which are present in every study by sample.

Random Sampling Errors. First, there are the random sampling errors or sampling fluctuations dealt with in the theory of probability and in the theory of sampling distributions. They are the outcome of the random sampling process, and sampling theory enables us to estimate them when we know their form of distribution. Random sampling errors have the advantageous property that they can be controlled by regulating the design and size of the sample. We have considerable theoretical and experimental knowledge of this type of error. Often, however, particularly in sampling survey studies, this is the smallest error in the collected data.

Systematic Errors. Apart from sampling fluctuations, errors also originate from the unreliability of human observers, either in direct observation or in other forms of measurement. Errors of measurement are usually much greater in biological, psychological, economic, or social investigations than in the physical sciences. Insofar as observational errors originate unconsciously, they may more or less follow the normal distribution of errors so that positive and negative deviations would tend to cancel increasingly as the number of observations increase. It is a mistake, however, to rely upon these errors' canceling one another. They may often possess not only a random element but also a bias. A special study needs to be carried out either by repeating the observations or measurements by the same observer or by more than one observer or by some other type of control and to compare the results. In making some observations, we are at times prone to dismiss as unessential conditions about which we think we know more or less. At times there may be justification for this attitude. It is good practice, however, to test the possibility of some circumstance as a cause by arranging the observations with respect to the circumstance. If the assumed cause is real, it is found that the errors of the observations display a regularity not found in chance errors. Wrong assumptions concerning the operation of some circumstance may bring about similar findings in calculations dealing with the results of observations. Errors of this type are called systematic errors.

Miscellaneous Inaccuracies. Contrasting sharply with random observational errors and sampling errors, inaccuracies may arise in a number of ways. The worst of these originate from such practices as false entries or entries by pure guess, deliberate violations of directions, or similar gross negligence. Other milder forms of inaccuracies, but nevertheless of substantial significance, need to be considered. Variations in kind and degree occur, dependent on problem, field, and method.

In making a house-to-house survey, it is obvious that much depends on the resourcefulness, skill, and reliability of the investigators. The kind of information obtained by asking questions on subjects that are poorly defined, or on matters of opinion, depends considerably on the form of questions asked. Sometimes the result of the inquiry is conditioned by the investigation itself, as, for instance, when the person interviewed may not have heard or thought of the subject before.

Much use of the questionnaire is made in collecting information from people who are not interested in statistics, and are often unwilling or unable to provide the information sought. People vary greatly in the trustworthiness of their returns, which are likely to be reasonably accurate only if the questions are few, well formulated, and easy to answer.

In complicated and difficult investigations, trained and experienced workers are necessary if the information collected is to be relied upon. For instance, special training and very complete directions as to how the forms are to be filled in are given to census enumerators.

Changed Conditions. Statistics extending over long periods are likely to be influenced by changes that may have taken place, particularly by new knowledge that may have altered the basis of classification or the ordering of things into classes. Improved systems of coverage and enumeration may render difficult comparisons of census data collected in different decades. Uniformity and precision in classifying can be achieved only if very complete and explicit definitions are given. Continuity is usually very significant in recorded statistics. In fact, at times the statistician may prefer an existing practice, so as to ensure continuity of records, to improved procedures. At any rate, if changes need to be made, he will insist on the collection of two sets of data, at least for some time—one under the old plan, the other under the new, so that continuity may be preserved.

This need for uniform conditions might be illustrated if an attempt were made to interpret the differences between the health status of men eligible for Army service in 1917 and in 1941. Such difficulties as the following would be likely to make any rigorous comparisons impossible: (1) the age groups are not identical, (2) the criteria for rejection are not the same, (3) changes in medical knowledge since 1917 have made possible the development of greatly improved techniques for identifying physical disabilities.

The valid interpretation of final statistical results requires a knowledge of the conditions surrounding the events recorded at the place and time of observation. For instance, there are many limitations on the use of physical-examination findings of selectees in World War II for drawing inferences concerning the general health status or the incidence of minor defects among the population (Ref. 14). The examinees at any induction station comprised a partly selected and widely variable sample of the male population at a specific time and place. The composition of the selectees chosen for examination was conditioned by (1) prevailing Selective Service policies with respect to deferments for dependency, (2) practices of the Armed Forces in regard to the acceptance of special groups, (3) the extent of differential screening of local boards, and (4) the number of men previously rejected who were sent up for re-examination. The comparison, for example, of those individuals who were rejected during the prewar period of Selective Service with those rejected at various periods during the war would require careful interpretation. The high rejection rates of the former do not necessarily imply a low level of national health.

Differing Types of Canvass. Deming (Ref. 6) enumerated and discussed 13 different factors that affect the usefulness of survey studies. This comprehensive and informative discussion includes additional types of errors or additional properties of errors not hitherto discussed. Only brief consideration can be given to these. Information is needed with respect to differences in results obtained from different kinds and degrees of canvass, such as mail, telephone, telegraphs, and interviews; also from different types of questionnaires. Different results are obtained by the different sponsoring agencies under whose auspices the survey study is carried out. For example, studies on income and work status yield different results when conducted by relief organizations than when conducted by a government agency. Because of this bias, government and private organizations have at times contracted with other agencies for the collection of data. Cohen (Ref. 2) reports an instance where in China one census, taken for poll-tax and military purposes, showed a population of 28,000,000. Another census over the same territory, taken this time for famine relief, returned a population of 105,000,000.

Changes in Population. There may be changes in the population in the interval between the time of collection of data and their processing. A sample may be more reliable than complete returns because of the shorter period required for collecting and processing. Because processing the data must commence at a certain date, replies received after this deadline are not included. The late reports may be biased. A sample study of these belated reports may at times determine whether bias is present. The comparison of two or more samples of the same sampling design or of subsamples within the main sample does not detect "systematic error" inherent in the methods. If two samples agree it may indicate not that they are devoid of bias but that their biases are similar.

Unrepresentative Date. Bias can occur from an unrepresentative choice of a date for a survey or a period to be covered. For instance, a passenger-traffic survey would not be representative if taken on or near a holiday date, nor would a school survey taken, say, the first week in June. Comparison of retail sales made in April, 1938, with those in April, 1937, gave spurious results, since the Easter holiday in 1937 came at the end of March, whereas in 1938 Easter occurred in the middle of April (Ref. 2).

In Processing. Processing errors may result from differences among workers in interpreting the wording of instructions, in editing, and in field work. Machine and tally errors need to be checked.

PLANNING THE INVESTIGATION

It should be noted that even if a 100 per cent sample were taken there would still remain errors of certain kinds enumerated here, such as bias of nonresponse (omissions), errors of response, late reports, errors originating in the tabulation plans, bias from unrepresentative dates or periods, changes taking place in the population before tabulations become available, and errors in interpretation. Furthermore, even if there is 100 per cent coverage, this is still a sample since at any other given time a new sample needs to be taken.

In the planning of an investigation by sample the research worker attempts to make the best possible effort to control the errors to which his study is susceptible. The distribution of his effort should be determined so that the greatest possible information will be obtained with the funds available. In fact, preliminary consideration of all the errors to which the projected study is liable largely determines whether or not the investigation should be carried out. Once a decision to proceed has been taken, the reduction in error will be dependent upon the wise distribution of funds such that the more significant sources of error will receive the most attention. Bias, consistency, and efficiency are dependent upon the system of sampling and estimation function used. The theoretical distinction between types of errors to be expected is clear. Sampling error and observational error of the random type are capable of statistical control. The amount of sampling error to be expected can be determined for each particular type of sampling design and size of sample. If the amount of error that can be tolerated is known, then it is an unwise use of resources to take a larger sample than is necessary.

Inaccurate instruments, the fallibility of human observers, defective techniques, biased methods of selecting data, and other such sources of systematic variation give errors which do not come within the scope of the classical theory of errors. These types of errors, therefore, need to be cared for largely by knowledge of and control of their sources. Systematic errors in the data may be larger than errors due to sampling.

Except for the random factors that might balance out, further increase in size of the sample does not increase the accuracy by eliminating systematic errors. Nor would they disappear if complete enumeration was resorted to.

It is an essential part of the sampling design to provide statistical controls for detecting and guarding against systematic types of errors. One way of doing this, for instance, in a sample survey is to collect two or more interpenetrating subsamples, which may be independent or partially linked together (Ref. 19). Such a simple control may not always suffice. It may be advisable, therefore, to arrange for the survey of the same sample, wholly or in part by two or more different workers. Just which sources of error are to receive the most attention will depend upon their importance in relation to the accuracy with which the study must be carried out in order to produce useful results with the funds available. This is the matter of the particular problem. Knowledge of the actual conditions and the types of systematic variation likely to arise in them, and how they may be eliminated or reduced when necessary, is basic.

The margin of error of the final estimate that can be tolerated if the conclusions drawn are to merit confidence must be considered in light of all kinds of errors to which the data are susceptible. The lack of accuracy and reliability in the data cannot, of course, be overcome by the subsequent statistical analysis that is applied. Thus, the task is first to secure data that are sufficiently precise for the purpose in hand and then to apply methods of analysis that make the best possible use of the information they contain.

PROCEDURES IN RANDOM SAMPLING

In random and in representative sampling, a fundamental assumption is that the sample is random. Upon the fulfillment of this assumption rests the validity of the application of most of the statistical analysis. The objective measurement of errors of estimation and the determination of the significance of the sample results are dependent on the hypothesis of the randomness of the sampling errors. It is, therefore, of interest and importance to note what solution, if any, the statistician formulates so that he can proceed with confidence in his analysis.

The information as to whether a sample is random is not available through examination of the properties of the sample itself. This shortcoming is illustrated by some of the hands which are obtained from dealing at random from a pack of cards, for instance, a hand containing 13 diamonds. The criterion, therefore, of a random sample has to be sought elsewhere, namely, in the process or method of selection. If a random method of selection can be developed, then a random sample can be simply defined as a sample which has been obtained by a random method.

The concept of a population comprised of aggregates of things or repeated events or phenomena is fundamental, since no collection of things can be thought of as random unless it in turn can be regarded as one of a set of such aggregates. Here it is assumed that a random set of objects means that the set was obtained by a random method. It is recalled that random sampling at the outset is designed to give every possible sample of given size an equal chance of being the actual sample. A requisite of a random sampling method is that it should be independent of the characteristics of the population under investigation. Since this definition of random selection refers to the specific character under study, it is evident that the random method in itself can not be thought of separately from the population the individuals of which are being chosen. A method might be random for one population and not for another. In fact, a method random for one characteristic of a population might not be so for another. Kendall (Ref. 11) illustrates this point by citing the problem of sampling in a planned town by taking every tenth house. This procedure might give a random sample, but if every tenth house should be a corner house, the sample might or might not be random, depending on the character that was being studied. It would probably preserve its randomness if, for instance, the study was to determine the proportion of inhabitants with blue eyes, but probably would lose it if the investigation was concerned with estimating the distribution of incomes.

No objective means is available for completely satisfying the requirement of independence between method of sampling and characteristic. Such means would require complete information about the population, which, if available, would of course render investigation unnecessary. Confidence in the fulfillment of independence must rest more or less in the actual state of our knowledge at the time on an a priori basis.

In practice, it is often essential to choose a sample random in relation to all properties of the population. This might appear to be an impossible task, since, as has been pointed out, it is in the very nature of the problem to sample the population according to at least one of its characteristics. This seeming predicament was removed by superimposing a new characteristic on the universe and sampling in accordance with it. The most useful characteristic that can be superimposed on an existent universe is that of *ordinal number*. If the universe can be enumerated, then the problem of random sampling becomes fundamentally that of discovering a series of random numbers.

The customary way is to number the universe in any practical manner, whether or not related to its properties, and then to look for a set of numbers so that they constitute a random aggregate from the possible ordinal numbers of the universe. Thus, rather than the requirement of determining in each case whether a sampling method is independent of the characteristics of the population, it became necessary only to construct a set of digits capable of giving a random sample of any size from any finite set of integers. Under such conditions it may be expected that the arrangement of digits in the sampling numbers will not be associated with the characteristics of the universe. Such was the principle upon which sets of "random sampling numbers" have been compiled.

Kendall (Ref. 11) specifies certain requirements, other than that of having been chosen at random, that a set of random sampling numbers must satisfy if it can be used for random sampling. Each digit in a set of N random sampling numbers is expected to occur in N/10 cases and each pair of digits to occur an equal number of times. He speaks of a set with such properties as locally random and gives four necessary tests, although they are not sufficient, to determine the existence of local randomness:

(1) The frequency test. Each digit should occur an approximately equal number of times.
(2) The serial test. There should be no tendency for a digit to be followed by any other digit.
(3) The poker test. There will be certain expectations to be satisfied for digits to be arranged in blocks of, say, five, four, three, and so on.
(4) The gap test. There are certain expectations to be satisfied with respect to the gaps occurring between the same digits in the series.

There are two sets of random sampling numbers in common use, Tippett's (Ref. 27) and Fisher and Yates's (Ref. 7). A third set has been published by Kendall and Smith (Ref. 12). Tippett compiled his set by drawing 41,600 digits at random from census reports and by combining in 4's to give 10,400 four-figured numbers. They have been subjected to a number of inquiries in which they have met the criteria of randomness used. Fisher and Yates's set of random numbers was constructed from the fifteenth and nineteenth digits of A. J. Thompson's 20-figure logarithmic table. The authors present tests of its randomness.

Each of the compilations is accompanied by a number of illustrations of its use. If, for instance, a random sample is wanted from a list or roster of names, the procedure would be as follows: First each sampling unit is numbered in any way, systematic or otherwise. The tables are then opened at random and starting at any point and proceeding in any direction, such as up or down the columns, along the rows, or by some other predetermined plan, a sufficient number of pairs of digits or othei combinations are taken to make up the predetermined size of the sample. Whenever the same number occurs twice or more it is simply ignored.

All numbers which exceed the total number of sampling units are also ignored.

Other methods of drawing random samples are used, such as using coins, dice, roulette wheels, or cards. Great care must be taken, however, to avoid bias in using such mechanical means. The human being has been shown to be especially incompetent to make a random selection. The problem of selecting a random sample has been greatly simplified by the preparation of tables of random sampling numbers. When the rules of the game are scrupulously observed, their use likely gives the best guarantee now available of obtaining a random sample.

A Comparative Experiment in Sampling Methods

In order to illustrate some of the principles underlying sampling procedures that have been discussed in this chapter, an experiment was carried out. Its findings are presented herewith.

We have a finite population consisting of 24,395 high school graduates whose ages were given as of the nearest birthday at the time of graduation. They have been classified according to sex and location of high school, as given in Table 45. The means and standard deviations in years for the total population and for each of the four subclasses are also recorded in Table 45.

TABLE 45

Ages of 1933–1944 High-School Graduates in Public Schools of Minnesota Classified according to Sex and Size of Locality*

Age	State as a whole	Outside 3 cities of first class		3 cities of first class	
		Boys	Girls	Boys	Girls
15	84	26	43	6	9
16	1,585	457	812	115	201
17	8,729	2,486	3,870	930	1,443
18	12,148	3,269	4,726	1,667	2,486
19	1,562	352	637	239	334
20	216	56	73	46	41
21	71	22	19	22	8
Total	24,395	6,668	10,180	3,025	4,522
Mean	17.59	17.56	17.53	17.74	17.68
S.D.	.7799	.7763	.7903	.7848	.7352

* State of Minnesota, Department of Education, Statistical Division, December, 1944.

We shall assume that we wish to estimate the age of high-school graduates by taking a sample of 1,000 from the total population of 24,395. We shall use three different methods of selecting the sample by assuming that each age group is (1) evenly distributed among the subclasses, (2) stratified proportionately to the sizes of subclasses, and (3) stratified proportionately to the products of the sizes and standard deviations of the subclasses.

First we shall describe the method of drawing the sample of 1,000 graduates from this population as a whole.

The first step was to assign a five-place number to each element of the population (see Table 46).

TABLE 46
Assignment of Random Sampling Numbers to the 24,395 Graduates of Table 45

Age	*Numbers*
15	00,001–00,084
16	00,085–01,669
17	01,670–10,398
18	10,399–22,546
19	22,547–24,108
20	24,109–24,324
21	24,325–24,395

The second step was to read Fisher and Yates's Table of Random Sampling Numbers (Ref. 7), page by page, first horizontally and then vertically. Each time five consecutive figures were read; they constituted a five-place number which was then referred to Table 46 to give the element an age score. Whenever a number larger than 24,395 was obtained, it was discarded. In this way, we formed a sample of 1,000 as indicated in Table 47.

TABLE 47
A Sample of 1000 Drawn by the Method of Random Sampling Numbers

Age	*State as a whole*
15	4
16	63
17	378
18	486
19	53
20	11
21	5
Total..........	1000

The final step was to stratify this sample of 1,000 according to the three methods enumerated above.

The first method, that of stratification with no restriction, was very simple. We simply split each age group into four subgroups as reported in Table 48.

In using the second method, that of stratification proportionate to the total number in the population in each of the four subclasses, we needed first to compute the proportions of the four subclasses. Let us

TABLE 48
STRATIFICATION OF THE SAMPLE OF 1000 GRADUATES WITH NO RESTRICTIONS

Age	Outside 3 cities of first class		3 cities of first class	
	Boys	Girls	Boys	Girls
15	1.00	1.00	1.00	1.00
16	15.75	15.75	15.75	15.75
17	94.50	94.50	94.50	94.50
18	121.50	121.50	121.50	121.50
19	13.25	13.25	13.25	13.25
20	2.75	2.75	2.75	2.75
21	1.25	1.25	1.25	1.25

denote by N_1 and N_2 the numbers of boys and girls respectively, outside the three cities; by N_3 and N_4, the numbers of boys and girls respectively, inside the three cities. Then we calculate:

$$\begin{aligned} N_1:N_2:N_3:N_4 &= 6{,}668:10{,}180:3{,}025:4{,}522 \\ &= \frac{6{,}668}{24{,}395}:\frac{10{,}180}{24{,}395}:\frac{3{,}025}{24{,}395}:\frac{4{,}522}{24{,}395} \\ &= .2733:.4173:.1240:.1854 \end{aligned}$$

Each age group was then split according to this ratio. The resultant stratification is reported in Table 49.

TABLE 49
STRATIFICATION OF THE SAMPLE OF 1000 GRADUATES ACCORDING TO PROPORTIONATE NUMBERS IN THE POPULATION STRATA

Age	Outside 3 cities of first class		3 cities of first class	
	Boys	Girls	Boys	Girls
15	1.09	1.67	0.50	0.74
16	17.22	26.29	7.81	11.68
17	103.31	157.74	46.87	70.08
18	132.82	202.81	60.26	90.10
19	14.48	22.12	6.57	9.83
20	3.01	4.59	1.36	2.04
21	1.37	2.09	0.62	0.93

In using the third method, that of stratification proportionate to the product of the numbers and standard deviations in the four subclasses, we also first have to compute the proportions of $N_i\sigma_i$'s of the four subclasses.

Let us assume that N_1, N_2, N_3, and N_4 have the same notation as used before. Denote by σ_1 and σ_2 the standard deviations of the ages of boys and girls respectively, outside the three cities; by σ_3 and σ_4, the respective standard deviations inside the three cities. Then we calculate:

$$\begin{aligned} N_1\sigma_1:N_2\sigma_2:N_3\sigma_3:N_4\sigma_4 &= 6{,}668(.7763):10{,}180(.7903):3{,}025(.7848):4{,}522(.7352) \\ &= 5{,}176:8{,}045:2{,}374:3{,}325 \\ &= \frac{5{,}176}{18{,}920}:\frac{8{,}045}{18{,}920}:\frac{2{,}374}{18{,}920}:\frac{3{,}325}{18{,}920} \\ &= .2736:.4252:.1255:.1757 \end{aligned}$$

Each age group was then split according to this ratio. The resulting stratification is reported in Table 50.

TABLE 50

STRATIFICATION OF THE SAMPLE OF 1000 GRADUATES PROPORTIONATE TO THE PRODUCT OF THE MEANS AND STANDARD DEVIATIONS IN THE POPULATION STRATA

Age	Outside 3 cities of first class		3 cities of first class	
	Boys	Girls	Boys	Girls
15	1.09	1.70	0.50	0.70
16	17.24	26.79	7.91	11.07
17	103.42	160.73	47.44	66.41
18	132.97	206.65	60.99	85.39
19	14.50	22.54	6.65	9.31
20	3.01	4.68	1.38	1.93
21	1.37	2.13	0.63	0.88

We then tested the goodness of fit for the three kinds of stratification by using the χ^2-criterion. Before doing this we needed to compute the theoretical expectations of frequencies for each age group in each subclass if we drew a sample of 1000 exactly representative of the parent population. The calculations of the theoretical expectations are reported in Table 51.

The test of the goodness of fit of the method of randomization without restrictions gave a value of $\chi_0^2 = 262.2836$. Referring to the χ^2 table with 18 degrees of freedom, we find that $P < .001$. Therefore, we conclude that this kind of stratification is not a good fit to the theoretical expectations.

The test of goodness of fit of the distribution of observed values from the method of stratification according to proportionate numbers and the theoretical distribution gave a value of $\chi_0^2 = 20.1521$. Referring to the χ^2 table with 18 degrees of freedom, we find that the corresponding value of $P > 30$. Therefore, we conclude that the stratification pro-

TABLE 51

CALCULATION OF THE THEORETICAL EXPECTATIONS OF FREQUENCIES FOR EACH AGE GROUP OF EACH SUBCLASS FOR A REPRESENTATIVE SAMPLE OF 1000 GRADUATES

		Age	F (Frequency of population)	Per Cent $\frac{F}{24,395}$	f_t (Theoretical frequency: % 1000)
Outside Three Cities	Boys	15	26	.00107	1.07
		16	457	.01873	18.73
		17	2,486	.10191	101.91
		18	3,269	.13400	134.00
		19	352	.01443	14.43
		20	56	.00230	2.30
		21	22	.00090	0.90
	Girls	15	43	.00176	1.76
		16	812	.03329	33.29
		17	3,870	.15864	158.64
		18	4,726	.19373	193.73
		19	637	.02611	26.11
		20	73	.00299	2.99
		21	19	.00078	0.78
Three Cities	Boys	15	6	.00025	0.25
		16	115	.00471	4.71
		17	930	.03812	38.12
		18	1,667	.06833	68.33
		19	239	.00980	9.80
		20	46	.00188	1.88
		21	22	.00090	0.90
	Girls	15	9	.00037	0.37
		16	201	.00824	8.24
		17	1,443	.05915	59.15
		18	2,486	.10191	101.91
		19	334	.01369	13.69
		20	41	.00168	1.68
		21	8	.00033	0.33
Total			24,395	1.00000	1000.00

portionate to subclass numbers in this case is a good fit to the theoretical expectations.

From the test of the goodness of fit for the method of stratification according to the product of the numbers and standard deviations in the sub classes, we found a $\chi_0^2 = 18.1743$. Referring to the χ^2 table with 18 degrees of freedom, we find that the corresponding value of $.50 > P > .30$. Therefore, we conclude that the stratification proportionate to the products of subclass numbers and standard deviations in this case is a good fit to the theoretical expectations. It is noted from Table 45 that the subclass standard deviations are all in the same magnitudes. Hence, the ratio of $N_1\sigma_1:N_2\sigma_2:N_3\sigma_3:N_4\sigma_4$ differs very little from the ratio of $N_1:N_2:N_3:N_4$. We do, however, note a reduction in χ^2 in taking into account the subclass standard deviations.

PROBLEMS

1. Work out a sampling design for securing data about the number of students enrolled in the several high-school subjects in your state.
2. Design a sampling survey for obtaining data concerning promotion policies for teachers in the elementary schools of your state.
3. Secure a representative sample of schools to engage in a cooperative experiment testing the relative efficacy of different curricular practices in secondary schools.
4. Design a sample survey for securing the best estimate of student enrollment in institutions of higher education in the United States; this information to be made available within a month after the opening of the institutions in the fall.
5. Set up a plan for a survey by sample of the attitude of the public toward Federal support of education to equalize educational opportunities.
6. Set up a sample of schools in your state which can be used recurrently for the collection of school statistics. Design the sample so that designated portions of the schools are taken out each year and new schools added so that no school carries an excessive burden.
7. Compare a method of sampling with the method of complete survey for a specified educational problem with respect to cost and time required to issue the results.
8. What recent developments have taken place in the techniques of questionnaire construction, in procedures in carrying out the interview, and in bringing about maximum returns from prospective respondents?
9. What methods have been developed to control error in the processing of survey data?
10. How can developments taking place in electrical and electronic equipment be applied to large sample surveys?
11. Suggest methods based on statistical and research principles which could be used for improving and standardizing procedures for collecting school statistics in your state.
12. Evaluate the sampling procedures used in Kinsey, Alfred C., Pomeroy, Wardell B., and Martin, Clyde E., *Sexual Behavior in the Human Male.* Philadelphia: W. B. Saunders Company, 1948.
13. Criticize the sampling methods used in the Revision of the Stanford-Binet Scale. See Marks, Eli S., "Sampling in the Revision of the Stanford-Binet Scale," *Psychological Bulletin,* Vol. 44 (1947), pp. 413–434.

14. Compare the relative efficiency of the three different sampling methods described in the text for estimating the ages of high-school graduates by calculating the mean and standard deviation for each method and comparing these estimates with the population values. Calculate the sampling errors for each method (See *Note* in Problem 15).

15. Specify methods of forming estimates and calculating sampling errors for each of the following sampling methods:

(a) Random sampling (no restrictions)
(b) Stratified sampling
(c) Cluster sampling
(d) Sub-sampling
(e) Stratification for two or more factors
(f) Balancing

Note: This problem should be postponed until the student has studied the techniques of analysis of variance and covariance.

References

1. Cochran, W. G., "The Use of the Analysis of Variance in Enumeration by Sampling," *Journal of the American Statistical Association*, Vol. 34 (1939), pp. 492–510.
2. Cohen, Jerome B., "The Misuse of Statistics," *Journal of the American Statistical Association*, Vol. 33 (1938), pp. 657–674.
3. Cornell, Francis G., "Sample Plan for a Survey of Higher Education Enrollment," *Journal of Experimental Education*, Vol. XV (1947).
4. Cowden, Dudley J., "An Application of Sequential Sampling to Testing Students," *Journal of the American Statistical Association*, Vol. 41 (1946), pp. 547–556.
5. Craig, A. T., "On the Mathematics of the Representative Method of Sampling," *Annals of Mathematical Statistics*, Vol. X (1939), pp. 26–34.
6. Deming, W. Edwards, "On Errors in Sampling," *American Sociological Review*, Vol. IX (1944), pp. 359–369.
7. Fisher, R. A., and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research.* Edinburgh: Oliver and Boyd, Ltd., 1943.
8. Jessen, R. J., "Statistical Investigation of a Sample Survey for Obtaining Farm Facts," *Iowa Agricultural Experiment Station Research Bulletin* 304 (1942).
9. Hansen, Morris H., and Hurwitz, William N., "On the Theory of Sampling from Finite Populations," *Annals of Mathematical Statistics*, Vol. XIV (1943), pp. 333–362.
10. Kendall, M. G., *The Advanced Theory of Statistics*, Vol. 1. London: Charles Griffin & Company, Ltd., 1945.
11. ———, and Smith, B. Babington, "Randomness and Random Sampling of Numbers," *Journal of the Royal Statistical Society*, Vol. 101 (1938), pp. 147–166.
12. ———, and ———, *Tables of Random Sampling Numbers, Tracts for Computers, No. 24.* London: Cambridge University Press, 1940.
13. Kermack, W. O., and McKendrick, A. G., "Tests for Randomness in a Series of Numerical Observations," *Proceedings of the Royal Society of Edinburgh*, Vol. LVII (1937), pp. 228–240.

14. Lew, Edward A., "Interpreting the Statistics of Medical Examination of Selectees," *Journal of the American Statistical Association*, Vol. 39 (1944), pp. 345–356.

15. Lindquist, E. F., *Statistical Analysis in Educational Research.* Boston: Houghton Mifflin Co., 1940, pp. 21–29.

16. Madhva, K. B., "Technique of Random Sampling," *Sankhya*, Vol. 4, Part 4 (1940), pp. 532–534.

17. Madow, Lillian H., "Systematic Sampling and Its Relation to Other Sampling Designs," *Journal of the American Statistical Association*, Vol. 41 (1946), pp. 204–217.

18. Madow, W. G., and Madow, L., "On the Theory of Systematic Sampling," *Annals of Mathematical Statistics*, Vol. XV (1944), pp. 1–24.

19. Mahalanobis, P. C., "On Large-Scale Sample Surveys," *Philosophical Transactions of the Royal Society (London)*, Series B, Biological Sciences No. 584, Vol. CCXXXI (1944), pp. 329–451.

20. McNemar, Quinn, "Sampling in Psychological Research," *Psychological Bulletin*, Vol. 37 (1940), pp. 331–365.

21. Neyman, Jerzy, "Contribution to the Theory of Sampling Human Populations," *Journal of the American Statistical Association*, Vol. 33 (1938), pp. 101–116.

22. ———, "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Vol. 97 (1934), pp. 558–625.

23. Snedecor, George W., "Design of Sampling Experiments in the Social Sciences," *Journal of Farm Economics*, Vol. XXI (1939), pp. 846–855.

24. Stephan, Frederick F., "Representative Sampling in Large-Scale Surveys," *Journal of the American Statistical Association*, Vol. 34 (1939), pp. 343–352.

25. ———, Deming, W. Edwards, and Hansen, Morris H., "The Sampling Procedure of the 1940 Population Census," *Journal of the American Statistical Association*, Vol. 35 (1940), pp. 615–630.

26. Sukhatme, P. V., "Contribution to the Theory of the Representative Method," *Supplement to Journal of the Royal Statistical Society*, Vol. II (1935), pp. 253–268.

27. Tippett, L. H. C., *Tables of Random Sampling Numbers, Tracts for Computers, No. 15.* London: Cambridge University Press, 1927.

28. Yates, F., "A Review of Recent Statistical Developments in Sampling and Sampling Surveys," *Journal of the Royal Statistical Society*, Vol. 109 (1946), pp. 12–30.

29. Yule, G. U., *The Statistical Study of Literary Vocabulary.* London: Cambridge University Press, 1944.

30. ———, and Kendall, M. G., *An Introduction to the Theory of Statistics.* London: Charles Griffin & Company, Ltd., 1937, pp. 332–335.

CHAPTER X

ANALYSIS OF VARIANCE AND COVARIANCE

The analysis-of-variance technique developed by R. A. Fisher and first reported in 1923 (Ref. 7) constitutes a method capable of analyzing the variation to which experimental and observational material is subject so that an assessment of the various components of variation can be made. Since its introduction, the analysis of variance has become more and more useful to large numbers of research workers in many fields. Fisher's technique is the only efficient one so far developed by which it is possible to differentiate the variation according to causes or groups of causes and to interpret the significance of a number of components simultaneously.

The modern advances in experimental and sampling designs have become possible through the development of exact tests of significance and of the analysis of variance. Without these tools, the assessment of the components of variation traceable to the sources specified by the experimental or sampling design would be a very involved and difficult enterprise. Fisher (Ref. 4) describes the analysis of variance as used in the analysis of experimental results as a simple arithmetical procedure for arranging and presenting the experimental results in a single compact table. This form of presentation shows both the structure of the experiment illustrated by the division of the number of degrees of freedom, and the relevant results arranged conveniently for the application of the necessary tests of significance.

The Analysis of Variation. Assume that we have a measure of a characteristic whose value is specified by the letter X. This value of X usually varies from one individual to another or for repeated measurements of the same individual. In general, the variation is due to a large number of different factors or causes. Of these factors some may be capable of identification and therefore may be called *assignable causes* of variation. However, there are usually numerous other causes which cannot be segregated because of our ignorance concerning them. These are spoken of as *chance causes.* As we gain in knowledge, more and more factors become assignable until the phenomenon can be completely explained if we can identify all the factors giving rise to the variation. The contribution of the known and unknown factors to the quantity X may be regarded, at least to a first approximation, as additive in character and may be represented symbolically thus:

$$X = a + b + c + \cdots + z \qquad (10.01)$$

where a, b, c, . . . denote the respective contributions of the known factors A, B, C, . . . , and z represents the residual or the portion of X attributable to chance or unknown factors. If, for instance, the factors A, B, C, . . . can be maintained under complete control, their respective contributions a, b, c, . . . will continue to be constant, whereas the fluctuations from unit to unit in X will be entirely attributable to the variation in z.

In experimental work various hypotheses may be advanced with respect to the effect of one or more factors, namely, A, B, C, . . . , and experimental designs are prepared to make the best determination of the presumed effects, a, b, c, ... The measures obtained of the presumed effects need to be tested with a view to determining their significance. If the measured effects are real, that is, traceable to the origin specified by the particular experimental design, the experimental results would be characterized as *heterogeneous* in variation. If, however, the variation presumably contributed by the several independent contributions of the factors A, B, C, . . . would be only of the order of magnitude of the effect assigned to the random sources of variation, the conclusion would be that the presumed effects were not real but attributable to random causes. The variation in the experimental material would then be spoken of as *homogeneous*. That is, in order for variation to be strictly homogeneous, it is purely random—caused by a multiplicity of minor independent factors, incapable of resolution into more elemental form and indistinguishable one from another.

Hence, the fundamental problem in studies of variation is to be able to differentiate the variation and to trace each contributing factor or group of factors to its source. Although an analysis of this kind is of special significance in experimental work, there are many situations in research work where differentiation of sources of variation in observational data is an essential part of the analysis. A general problem is that of determining whether two or more samples may be regarded as random samples from the same homogeneous population.

An Application of Analysis of Variation. We shall illustrate the main ideas of the above discussion by presenting an example. Let us take the data recorded in Table 52 which represent the mental ages in months of 6 samples of 6 pupils each, each randomly chosen from the same grade in 6 different urban schools. Suppose that the data are required to answer the question: Is there evidence that the mental ages of the pupils are the same for the same grade in the 6 schools?

The variability in the mental ages of the pupils from the same school is so considerable that it would be hard to reach a conclusion on the point at issue from a mere inspection of the data in the table. Diagram 5 brings out the situation more clearly; but even after examining it can we say that the differences in the means are significant? It is at this point

TABLE 52

Mental Ages of 6 Random Samples of 6 Pupils Each from 6 Different Schools

Individual	Mental ages in schools					
	1	2	3	4	5	6
1	158	156	160	159	164	153
2	157	155	158	155	163	148
3	153	154	156	148	162	145
4	151	153	155	147	160	144
5	144	151	150	146	154	144
6	143	149	145	145	151	136
Mean	151	153	154	150	159	145

Grand mean = 152

that statistical theory can give assistance by determining how much consideration should be given to the apparent differences in means, which are hard to discern because of the residual fluctuation, z, due to chance causes. Specifically, the question is: What is the probability that the observed differences in the mean values of the 6 schools might have arisen simply through random sampling errors?

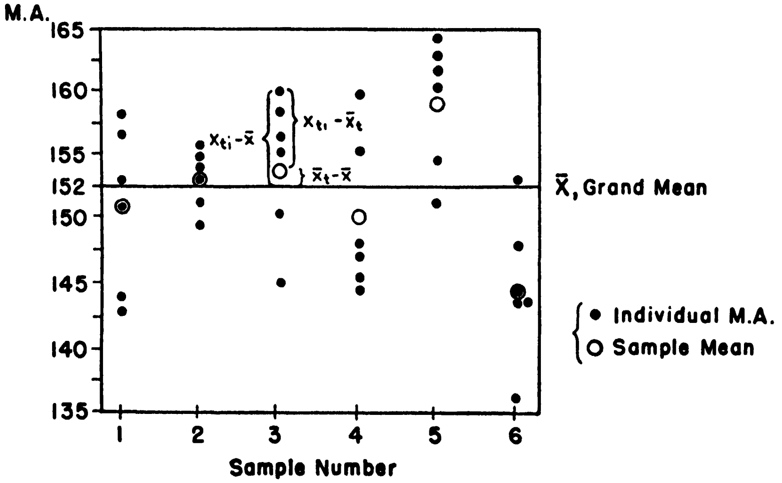


Figure 5. The components of variation in the mental ages of 36 pupils.

Statistical means enable us to make the calculation of the probability value. Since the form of the statistical test to be described later depends to a considerable extent on the nature of the random variation represented by z, it may be well to point out here the following assumptions: the random elements, in successive observations, are independent of each other and of the values of the assignable factors, if these exist; and they are normally distributed about zero with equal standard deviation.

We shall briefly describe the model the statistician sets up for the description of the situation discussed here. If the X of Equation (10.01) represents the mental age of a single individual, then a on the right-hand side may be considered as a general average or mean of the individuals in all the samples, and b as a contribution—positive or negative—associated with a particular sample. If there are changes in mental age from sample to sample which affect X, then the values of b, namely, b_1, b_2, . . . , b_6 for the 6 samples, will differ; if there are no such changes, then

$$b_1 = b_2 = \cdots = b_6 = 0 \qquad (10.02)$$

The random or residual variations, z, among the mental ages of individuals from the same school obscure the real situation about the true value as estimated from the sample. Hence, it is not possible to take the difference between the observed sample mean and the grand mean as equal to b_t ($t = 1, 2, \cdots, 6$). Therefore, it becomes necessary to answer the question: Taking into account the observed variation among mental ages of individuals in the same school sample, what is the probability that the 6 obtained sample means would differ so much among themselves because of random sampling fluctuations if, in fact, Equation (10.02) were true?

The method used by the statistician to solve this problem is outlined below.

Let X_{ti} be the mental age score of the ith individual in the tth sample; $i = 1, 2, \cdots, 6$; also $t = 1, 2, \cdots, 6$. $\bar{X}_t$ is the mean of the observations in the tth sample and $\bar{X}$ is the grand mean of the 36 observations. As illustrated for one individual from the third sample in Diagram 5, the mental age score of the ith individual in the tth sample may be considered as the sum of three components. Thus:

$$X_{ti} = \bar{X} + (\bar{X}_t - \bar{X}) + (X_{ti} - \bar{X}_t) \qquad (10.03)$$

For example, the mental-age score (164) of the first individual in the sample from the fifth school is equal to $152 + (159 - 152) + (164 - 159)$. Referring to Equation (10.01), $\bar{X}$ may be considered as an estimate of a; $(\bar{X}_t - \bar{X})$ as an estimate of b_t; and $(X_{ti} - \bar{X}_t)$ of the residual variation z_{ti}. These are estimates because we have observations only from a random sample from each of the schools.

The significance of the difference $\bar{X}_t - \bar{X}$ ($t = 1, 2, \cdots, 6$) or the acceptance of the hypothesis represented by Equation (10.02) is based on the magnitude of the components $\bar{X}_t - \bar{X}$ compared with $X_{ti} - \bar{X}_t$. A precise statistical test of the significance involves the use of the following identity:

$$\left.\begin{aligned}\sum_t \sum_i (X_{ti} - \bar{X})^2 &= \sum_t \sum_i [(X_{ti} - \bar{X}_t) + (\bar{X}_t - \bar{X})]^2 \\ &= \sum_t \sum_i (X_{ti} - \bar{X}_t)^2 + \sum_t \sum_i (\bar{X}_t - \bar{X})^2 \\ &\qquad + 2 \sum_t \sum_i (X_{ti} - \bar{X}_t)(\bar{X}_t - \bar{X}) \\ &= \sum_t \sum_i (\bar{X}_t - \bar{X})^2 + \sum_t \sum_i (X_{ti} - \bar{X}_t)^2\end{aligned}\right] \qquad (10.04)$$

since the product term will vanish because $\sum_i (X_{ti} - \bar{X}_t) = 0$.

Before the magnitude of the two components can be compared, they must be divided by the quantities known as the *number of degrees of freedom*, which are r and $N - q$, respectively, where r is the number of relations used to define the hypothesis, that is, 5 in this problem. There are 6 independent values; therefore, $q = 6$. If the hypothesis tested is true, then 5 relations hold among the 6 parameters, namely, $b_1 = 0$, $b_2 = 0$, $b_3 = 0$, $b_4 = 0$, $b_5 = 0$. Thus, $r = 5$; $N - q = 36 - 6 = 30$.

The criterion is

$$F = \frac{\dfrac{\sum_{ti} (\bar{X}_t - \bar{X})^2}{r}}{\dfrac{\sum_{ti} (X_{ti} - \bar{X}_t)^2}{N - q}} \qquad (10.05)$$

$$z = \tfrac{1}{2} \log_e \frac{\dfrac{\sum_{ti} (\bar{X}_t - \bar{X})^2}{r}}{\dfrac{\sum_{ti} (X_{ti} - \bar{X}_t)^2}{N - q}} \qquad (10.06)$$

Using the Tables of F or z, respectively, we obtain the 5 per cent and 1 per cent levels of significance against which the obtained value of F or z is checked.

The numerical solution for the example is carried out as follows:

First, it is convenient to reduce the values in Table 52 by subtracting 150 from each value obtaining the following:

Individual	1	2	3	4	5	6	Total
1	8	6	10	9	14	3	
2	7	5	8	5	13	− *2*	
3	3	4	6	*−2*	12	− *5*	
4	1	3	5	*−3*	10	− *6*	
5	*−6*	1	0	*−4*	4	− *6*	
6	*−7*	*−1*	*−5*	*−5*	1	*−14*	
Total	6	18	24	0	54	*−30*	72

We then calculate the "within schools" sum of squares, that is, the sum of the squares of the deviations of the mental ages of the individuals in a school (sample) about their school means, as follows:

$$\sum_t \sum_i (X_{ti} - \bar{X}_t)^2 = (8^2 + 7^2 + 3^2 + \cdots + [-14]^2) - \left(\frac{6^2 + 18^2 + 24^2 + 0^2 + 54^2 + (-30)^2}{6}\right)$$
$$= 1638 - 792$$
$$= 846$$

We next calculate the "between schools" sum of squares, that is, the sum of squares of deviations of the school means about the grand mean, as follows:

$$\sum_t \sum_i (\bar{X}_t - \bar{X}) = \frac{6^2 + 18^2 + 24^2 + 0^2 + 54^2 + (-30)^2}{6} - \frac{(72)^2}{36}$$
$$= 792 - 144$$
$$= 648$$

The total sum of squares, that is, the sum of squares of the deviations of the 36 individual mental ages from their grand mean, is obtained as follows:

$$\sum_t \sum_i (X_{ti} - \bar{X}) = (8^2 + 7^2 + 3^2 + \cdots [-14]^2) - \frac{(72)^2}{36}$$
$$= 1638 - 144$$
$$= 1494$$

The respective sums of squares with the appropriate number of degrees of freedom are recorded in the customary analysis-of-variance table,

TABLE 53
ANALYSIS OF VARIANCE OF THE MENTAL AGES OF THE 36 PUPILS IN 6 DIFFERENT SCHOOLS

Source of variation	d.f.	Sums of squares	Mean square	F	Hypothesis
Between schools	5	648	129.6	4.6	Rejected
Within schools	30	846	28.2		
Total	35	1494			

Table 53. The values under the column heading "mean square" are obtained by dividing the sum of squares in each row by the corresponding number of degrees. By applying Formula (10.05), we obtain as the

observed value of the criterion, F_0:

$$F_0 = \frac{129.6}{28.2}$$
$$= 4.6$$

We then enter the F-table with $n_1 = 5$, $n_2 = 30$, and find that $F_{.01} = 3.7$. Since our obtained value, 4.6, is greater than 3.7, we may conclude that there is a significant difference between the mean mental ages of the schools. We may also say that the null hypothesis under test, that is, the hypothesis stated in Equation (10.02), is rejected.

Process of the Analysis of Variance. As has been observed in the example above, the actual process in the analysis of variance consists in breaking up the total sum of squares of deviations of the observations from the grand mean into independent portions assigned to certain factors. The structure of these component parts, usually determined by the design of experiment, is specified by the number of degrees of freedom or by the number of independent comparisons, which, like the corresponding sums of squares, are additive in character. Therefore, the method is equally valid for small and large samples.

Analysis of Covariance. Another useful extension of the general analysis-of-variance method is the analysis of covariance, also developed by Fisher. In this analysis, the process consists in breaking down the sum of products of deviations of any two variates from their means and assigning the respective components to specified sources. One of the most useful applications of the covariance method is in sorting out the covariance effects, particularly in experimentation. This operation makes it possible to increase the precision of an experiment by the elimination of causes of variation in some cases not controlled or controllable by the experimental design.

Superiority of Analysis of Variance to the Traditional Biometric Method. While in experimentation the special value of the analysis of variance is manifest, it has many other applications in dealing with observational material. The efficiency of its use in testing if a group of samples may be regarded as having come from the same homogeneous population is clearly illustrated by comparison with the traditional biometric method used for such purposes. In the latter it is customary to calculate independently a standard error for each of the possible comparisons of the means of several samples. The labor involved in this procedure is not its only objection. The chief objection is that in many cases the obtained estimates of standard errors may not differ beyond merely sampling errors. In such cases it may be concluded that the larger part of the observed differences is attributable to random sampling errors, and that a more accurate as well as much less complicated analysis would result by pooling the sums of squares of deviations from the different means and by applying the combined estimate in the test of significance. This change introduced by the analysis-of-variance method serves to provide an exact test of the null hypothesis and hence is used habitually by the modern research worker. Thus the method makes use of the relevant information contained in the data, since it takes into account the sampling distribution of statistics of the same kind.

The foregoing discussion serves to give a general account of the main ideas underlying the analysis of variation. Accompanied by the illustrative example, this discussion should be suitable as an introduction for the reader to the application of the analysis of variance to the simpler problems. Probably, however, the research worker will profit from a more complete and rigorous study of the statistical principles underlying such a powerful tool as the analysis of variance and covariance. It is frequently observed that the formulation of a problem in statistical terms, which requires an orderly arrangement of the known results and an awareness of the assumptions and how they may be tested, assists in making clear the essential features of a problem hitherto not clearly visualized.

Before a number of practical applications of the method of analysis of variance and covariance are demonstrated, the next section will present the systematic formulation and solution of the problems underlying these methods. This section may be omitted by the reader not interested in the mathematical developments. He can proceed directly to the practical problems in Chapter XI.

Mathematical Foundations of Analysis of Variance and Covariance

Mathematical Ratification. 1. Suppose we have a normal distribution with mean μ and standard deviation σ. It is well known that if we pick independently all the possible samples of size n from this population and denote the random effects for each sample by

$$z_t = Y_t - \mu(t = 1, \cdots, n) \tag{1.01}$$

then the mean value of z_t will be normally distributed with mean 0 and standard deviation $\sigma/\sqrt{n}$. So we may define, in this case, the maximum likelihood estimate of the variance, σ^2, of the population as

$$\sigma^2 = n\sigma^2_{\bar{z}_t} \tag{1.02}$$

where $\sigma^2_{\bar{z}_t}$ is the variance of sampling means of the random effects.

The analysis-of-variance method consists in the breaking up of the total variance into independent parts which can produce independently the maximum-likelihood estimates of σ^2 due to the random effects alone. For instance, if we have p groups which are chosen by a certain criterion, then we immediately know in advance that these groups are more or less heterogeneous with respect to their corresponding means. However, we pretend to assume that they are randomly chosen from the whole population in presenting the mathematical formulation as follows:

$$z_{st} = Y_{st} - \mu - \alpha_s \qquad (s = 1, \cdots, p; t = 1, \cdots, n) \tag{1.03}$$

where z_{st} is the random effect; Y_{st} is an observation of the *t*th individual in the *s*th group; μ is the population mean; and α_s is the deviation from the population mean for the *s*th group. By the maximum-likelihood method we can easily get two independent estimates of σ^2 from our sample:

$$\sigma_1^2 = \frac{1}{p(n-1)} \sum_s \sum_t (Y_{st} - \bar{Y}_s)^2 \tag{1.04}$$

$$\sigma_2^2 = \frac{n}{p-1} \sum_s (\bar{Y}_s - \bar{Y}.)^2 \tag{1.05}$$

where $$\bar{Y}_s = \frac{\sum_t Y_{st}}{n}; \qquad \bar{Y}. = \frac{\sum_s \sum_t Y_{st}}{pn}.$$

By using Fisher's *z*-test or the variance ratio *F*, we can immediately determine whether or not these two variances are of the same magnitude.

Ordinarily, we are interested only in knowing if these groups have the same means. So we often make the test on the basis of σ_1^2, which is called the *variance of "within."* However, the result of significance of the variance σ_2^2, which is called the *variance of "between,"* implies three alternative explanations. These groups have

(1) Different means and different variabilities.
(2) The same mean and different variabilities.
(3) Different means and the same variability.

Therefore, if we wish to rule out the first two explanations, we have to test the hypothesis $\sigma_s = \sigma$ for these groups. This may be done by using the L_1-criterion.[1]

The same mathematical approach can be applied to the problems of more than one classification. In this case, we have independent estimates of σ^2 due to the interactions in addition to those due to the main factors.[2]

From the above, we wish to present assumptions which should be fulfilled in the analysis of variance:[3]

(1) The population distribution should be normal. This assumption, however, is not especially important. Eden and Yates (Ref. 2) showed

[1] For the method of using the L_1-criterion, see page 82.

[2] For a detailed consideration of these interactions, see Refs. 13 and 14 of Chapter XIII.

[3] For assumptions underlying the analysis of variance, see Ref. 3; for a discussion of the consequences when any assumption is not satisfied, see Ref. 1.

that even with a population departing considerably from normality, the effectiveness of the z-distribution still held. The normality and independence of the random elements in successive observations has been pointed out on page 212.

(2) All groups of a certain criterion or of the combination of more than one criterion should be randomly chosen from the subpopulation having the same criterion or having the same combination of more than one criterion. For instance, if we wish to select two groups in a school population, one of the third grade and the other of the fourth grade, we must choose randomly from the respective subpopulations. This assumption is the keystone of the analysis-of-variance technique. Failure to fulfill this assumption gives biased results.

(3) The subgroups under investigation should have the same variability. We should test this assumption before we run the analysis of variance. Otherwise, a false interpretation of the results may follow.

Maximum-Likelihood Solution of Analysis-of-Variance Problems. *With One Classification.* 2. Before we develop a general solution of the problems with any number of classifications, we start with the derivation of the solution for the problems with only one classification. The frequencies in different subclasses will always be assumed to be equal. We denote by Y_{st} the score obtained by the tth individual in the sth subclass. The basic assumption in the analysis of variance is that we may write

$$Y_{st} = M + A_s + z_{st} \tag{2.01}$$

where $s = 1, \cdots, p$; $t = 1, \cdots, n$; p denotes the number of subclasses; n denotes the number of individuals in each subclass; M is defined as the general mean; A_s is the deviation due to the sth subclass; and z_{st} represents the random effect for the tth individual in the sth subclass. To minimize the variance of z_{st} by using the maximum-likelihood method, we first write

$$\chi^2 = \sum_s \sum_t (Y_{st} - M - A_s)^2 + \lambda \sum_s A_s \tag{2.02}$$

where

$$\sum_s A_s = 0 \tag{2.03}$$

which is a restriction imposed on (2.01); and λ is an undetermined multiplier of Lagrange. Differentiating χ^2 partially with respect to M and A_s, setting the resulting equations equal to zero, and solving, we obtain

$$M = \frac{1}{N} \sum_s \sum_t Y_{st} = \bar{Y}. \qquad (N = pn) \tag{2.04}$$

$$A_s = \frac{1}{n} \sum_t Y_{st} - M - \frac{\lambda}{2n} = \bar{Y}_s - \bar{Y}. - \frac{\lambda}{2n} \tag{2.05}$$

From (2.03) and (2.05), we obtain

$$(2.06)\qquad \sum_s A_s = \sum_s \bar{Y}_s - p\bar{Y}. - \frac{\lambda}{2n} = 0$$

which reduces to

$$(2.07)\qquad \lambda = 0$$

By the method of elimination, we have

$$(2.08)\qquad \chi_a^2 = \sum_s \sum_t (Y_{st} - \bar{Y}_s)^2 = \sum_s \sum_t Y_{st}^2 - n \sum \bar{Y}_s^2$$

The hypothesis we wish to test is

$$(2.09)\qquad H_0 : A_s = 0$$

that is, the hypothesis that there is no significant difference between these subclasses. Assuming that H_0 is true, we have from (2.02),

$$(2.10)\qquad \chi^2 = \sum_s \sum_t (Y_{st} - M)^2$$

Minimizing χ^2 with respect to M, we obtain

$$(2.11)\qquad M = \bar{Y}.$$

Substituting this value into Equation (2.10), we obtain the relative minimum value $\chi_{r_0}^2$:

$$(2.12)\qquad \chi_{r_0}^2 = \sum_s \sum_t (Y_{st} - \bar{Y}.)^2 = \sum_s \sum_t Y_{st}^2 - N\bar{Y}_.^2 = \chi_a^2 + n \sum_s \bar{Y}_s^2 - N\bar{Y}_.^2$$
$$= \chi_a^2 + \chi_0^2$$

The additive property of the sum of squares is readily demonstrated in (2.12). All the results obtained may be summarized as in Table 54:

TABLE 54
Analysis of Variance for a Single Classification

Source of variation	D.F.	Sum of squares
Within subclasses	$N - p$	$\chi_a{}^2$
Between subclasses	$p - 1$	$\chi_0{}^2$
Total	$N - 1$	$\sum_s \sum_t Y_{st}{}^2 - N\bar{Y}_.{}^2$

With Two Classifications. 3. Now we shall work out the equations with two classifications—say column and row. We denote by $Y_{s_1s_2t}$ the score obtained by the tth individual in the s_1th column and the s_2th row.

The basic assumption in the analysis of variance is that we may write

$$Y_{s_1s_2t} = M + A_{s_1} + B_{s_2} + I_{s_1s_2} + z_{s_1s_2t} \tag{3.01}$$

subject to the following restrictions:

$$\sum_{s_1} A_{s_1} = 0 \tag{3.02}$$

$$\sum_{s_2} B_{s_2} = 0 \tag{3.03}$$

$$\sum_{s_1}\sum_{s_2} I_{s_1s_2} = 0 \tag{3.04}$$

where $s_1 = 1, \cdots, p_1$; $s_2 = 1, \cdots, p_2$; $t = 1, \cdots, n$; p_1 denotes the number of columns; p_2 denotes the number of rows; n denotes the number of individuals in each subclass; M is defined as the general mean; A_{s_1} is the deviation due to the s_1th column; B_{s_2} is the deviation due to the s_2th row; $I_{s_1s_2}$ represents the influence of the interaction between column and row; and $z_{s_1s_2t}$ represents the random effects. To obtain the solution, we first write

$$\chi^2 = \sum_{s_1}\sum_{s_2}\sum_{t} (Y_{s_1s_2t} - M - A_{s_1} - B_{s_2} - I_{s_1s_2})^2 + \alpha_1 \sum_{s_1} A_{s_1} + \alpha_2 \sum_{s_2} B_{s_2} + \alpha_3 \sum_{s_1}\sum_{s_2} I_{s_1s_2} \tag{3.05}$$

where α_1, α_2, and α_3 are the undetermined multipliers of Lagrange. Minimizing χ^2 with respect to M, A_{s_1}, B_{s_2}, and $I_{s_1s_2}$, we obtain

$$M = \frac{1}{N}\sum_{s_1}\sum_{s_2}\sum_{t} Y_{s_1s_2t} = \bar{Y}..(N = p_1p_2n) \tag{3.06}$$

$$A_{s_1} = \frac{1}{p_2n}\sum_{s_2}\sum_{t} Y_{s_1s_2t} - M - \frac{\sum_{s_2} I_{s_1s_2}}{p_2} - \frac{\alpha_1}{2np_2} \tag{3.07}$$

$$= \bar{Y}_{s_1}. - \bar{Y}.. - \frac{\sum_{s_2} I_{s_1s_2}}{p_2} - \frac{\alpha_1}{2np_2}$$

$$B_{s_2} = \frac{1}{p_1n}\sum_{s_1}\sum_{t} Y_{s_1s_2t} - M - \frac{\sum_{s_1} I_{s_1s_2}}{p_1} - \frac{\alpha_2}{2np_1} \tag{3.08}$$

$$= \bar{Y}._{s_2} - \bar{Y}.. - \frac{\sum_{s_1} I_{s_1s_2}}{p_1} - \frac{\alpha_2}{2np_1}$$

$$I_{s_1s_2} = \frac{1}{n}\sum_{t} Y_{s_1s_2t} - M - A_{s_1} - B_{s_2} - \frac{\alpha_3}{2n} \tag{3.09}$$

$$= \bar{Y}_{s_1s_2} - \bar{Y}.. - A_{s_1} - B_{s_2} - \frac{\alpha_3}{2n}$$

From (3.02) and (3.07), we have

$$\sum_{s_1} A_{s_1} = \sum_{s_1} \bar{Y}_{s_1\cdot} - p_1\bar{Y}_{\cdot\cdot} - \frac{\sum_{s_1}\sum_{s_2} I_{s_1s_2}}{p_2} - \frac{\alpha_1}{2np_2} = 0 \tag{3.10}$$

which immediately reduces to

$$\alpha_1 = 0 \tag{3.11}$$

Similarly,

$$\alpha_2 = \alpha_3 = 0 \tag{3.12}$$

By the method of elimination, we get

$$\chi_a^2 = \sum_{s_1}\sum_{s_2}\sum_{t} (Y_{s_1s_2t} - \bar{Y}_{s_1s_2})^2 = \sum_{s_1}\sum_{s_2}\sum_{t} Y^2_{s_1s_2t} - n\sum_{s_1}\sum_{s_2} \bar{Y}^2_{s_1s_2} \tag{3.13}$$

The hypothesis we wish to test first is

$$H_1 : I_{s_1s_2} = 0 \tag{3.14}$$

that is, the hypothesis that there is no influence of the interaction between column and row. Assuming that H_1 is true, we have, from (3.05),

$$\chi^2 = \sum_{s_1}\sum_{s_2}\sum_{t} (Y_{s_1s_2t} - M - A_{s_1} - B_{s_2})^2 + \beta_1 \sum_{s_1} A_{s_1} + \beta_2 \sum_{s_2} B_{s_2} \tag{3.15}$$

where β_1 and β_2 are the undetermined multipliers of Lagrange. Minimizing χ^2 with respect to M, A_{s_1}, and B_{s_2}, we obtain

$$M = \bar{Y}_{\cdot\cdot} \tag{3.16}$$

$$A_{s_1} = \bar{Y}_{s_1\cdot} - \bar{Y}_{\cdot\cdot} - \frac{\beta_1}{2p_2n} \tag{3.17}$$

$$B_{s_2} = \bar{Y}_{\cdot s_2} - \bar{Y}_{\cdot\cdot} - \frac{\beta_2}{2p_1n} \tag{3.18}$$

where

$$\beta_1 = \beta_2 = 0 \tag{3.19}$$

Substituting these values in (3.15) and simplifying, we obtain the relative minimum value $\chi^2_{r_1}$:

$$\begin{aligned} \chi^2_{r_1} &= \sum_{s_1}\sum_{s_2}\sum_{t} (Y_{s_1s_2t} - \bar{Y}_{s_1\cdot} - \bar{Y}_{\cdot s_2} + \bar{Y}_{\cdot\cdot})^2 \\ &= \chi_a^2 + \sum_{s_1}\sum_{s_2}\sum_{t} (\bar{Y}_{s_1s_2} - \bar{Y}_{s_1\cdot} - \bar{Y}_{\cdot s_2} + \bar{Y}_{\cdot\cdot})^2 \\ &= \chi_a^2 + n\sum_{s_1}\sum_{s_2} \bar{Y}_{s_1s_2} - p_2n\sum_{s_1} \bar{Y}_{s_1\cdot} - p_1n\sum_{s_2} \bar{Y}_{\cdot s_2} + N\bar{Y}_{\cdot\cdot} \\ &= \chi_a^2 + \chi_1^2 \end{aligned} \tag{3.20}$$

Then we may test the relative hypothesis on the basis of $\chi^2_{r_1}$:

$$H_{01}: A_{s_1} = 0 \tag{3.21}$$

that is, the hypothesis that there is no significant difference between columns. Assuming that H_{01} is true, we may write

$$\chi^2 = \sum_{s_1} \sum_{s_2} \sum_{t} (Y_{s_1 s_2 t} - M - B_{s_2})^2 + \gamma \sum_{s_2} B_{s_2} \tag{3.22}$$

where γ is the undetermined multiplier of Lagrange. Minimizing χ^2 with respect to M and B_{s_2}, we have

$$M = \bar{Y}.. \tag{3.23}$$

$$B_s = \bar{Y}._{s_2} - \bar{Y}.. - \frac{\gamma}{2p_1 n} \tag{3.24}$$

where

$$\gamma = 0 \tag{3.25}$$

Substituting these values into (3.22) and simplifying, we obtain

$$\begin{aligned} \chi^2_{r_{01}} &= \sum_{s_1} \sum_{s_2} \sum_{t} (Y_{s_1 s_2 t} - \bar{Y}._{s_2})^2 = \chi^2_a + \chi^2_1 + p_2 n \sum_{s_1} \bar{Y}^2_{s_1}. - N\bar{Y}^2_{\cdot}. \\ &= \chi^2_a + \chi^2_1 + \chi^2_{01} \end{aligned} \tag{3.26}$$

Finally, we may test the relative hypothesis on the basis of $\chi^2_{r_{01}}$:

$$H_{02}: B_{s_2} = 0 \tag{3.27}$$

that is, the hypothesis that there is no significant difference between rows. Assuming that H_{02} is true and proceeding as before, we obtain

$$\begin{aligned} \chi^2_{r_{02}} &= \sum_{s_1} \sum_{s_2} \sum_{t} (Y_{s_1 s_2 t} - \bar{Y}..)^2 = \chi^2_a + \chi^2_1 + \chi^2_{01} + p_1 n \sum_{s_2} \bar{Y}^2_{\cdot s_2} \\ &\qquad - N\bar{Y}^2_{\cdot}. \\ &= \chi^2_a + \chi^2_1 + \chi^2_{01} + \chi^2_{02} \end{aligned} \tag{3.28}$$

From (3.28), the additive property of the sum of squares is again clearly demonstrated. It is also noted, in the case of equal frequencies in sub-

TABLE 55

ANALYSIS OF VARIANCE FOR THE PROBLEMS OF DOUBLE CLASSIFICATION

Source of variation	D.F.	Sum of squares
Within................	$N - p_1 p_2$	χ_a^2
Column × row.........	$(p_1 - 1)(p_2 - 1)$	χ_1^2
Column...............	$p_1 - 1$	χ_{01}^2
Row..................	$p_2 - 1$	χ_{02}^2
Total	$N - 1$	$\sum_{s_1} \sum_{s_2} \sum_{t} Y^2_{s_1 s_2 t} - N\bar{Y}^2..$

classes, that there is only one answer for each hypothesis tested, no matter what the order of testing may be. All the results obtained may be summarized as in Table 55.

4. In general, if we have a problem of k classifications, the mathematical expression of the score made by the tth individual in the s_1th group of classification A, the s_2th group of classification B, . . . , and the s_kth group of classification R is as follows:

$$(4.01)\quad Y_{s_1s_2,\ldots,s_kt} = M + A_{s_1} + B_{s_2} + \cdots + R_{s_k} + I_{s_1s_2} + \cdots + I_{s_{k-1}s_k} + I_{s_1s_2s_3} + \cdots + I_{s_{k-2}s_{k-1}s_k} + \cdots + I_{s_1s_2,\cdots,s_k} + z_{s_1s_2,\cdots,s_kt}$$

where $s_1 = 1, \cdots, p_1$; $s_2 = 1, \cdots, p_2$; $\cdots$; $s_k = 1, \cdots, p_k$; p_1 denotes the number of groups in classification A; p_2 denotes the number of groups in classification B; . . . ; p_k denotes the number of groups of classification R; M is the grand mean; A, B, . . . , and R are the measures of the main effects with respect to their own subscripts; I's are the measures of the interactions with respect to their own subscripts; and $z_{s_1s_2,\cdots,s_kt}$ is the error. The solutions for the sum of squares of each source of variation are as follows:

1. Within:

$$(4.02)\quad \underbrace{\sum_{s_1}, \cdots, \sum_{s_k}}_{k\text{-fold}} \sum_{t} Y^2 - n \underbrace{\sum_{s_1}, \cdots, \sum_{s_k}}_{k\text{-fold}} \bar{Y}^2_{\underbrace{s_1,\cdots,s_k}_{\substack{k\\ \text{subscripts}}}}$$

2. Interactions and main effects:

$$(4.03)\quad \delta_1 \underbrace{\sum_{s_i}, \cdots, \sum_{s_j}}_{\substack{r\text{-fold}\\ i<\cdots<j}} \bar{Y}^2_{\underbrace{s_i,\cdots,s_j}_{\substack{r\\ \text{subscripts}\\ i<\cdots<j}}} - \mathop{S}\limits_{1}^{r}\Big(\delta_2 \underbrace{\sum_{s_i}, \cdots, \sum_{s}}_{\substack{r\text{-1 fold}\\ i<\cdots<}} \bar{Y}^2_{\underbrace{s_i,\cdots,s_j}_{\substack{r\text{-1}\\ \text{subscripts}\\ i<\cdots<j}}}\Big)$$
$$+ \mathop{S}\limits_{1}^{C_2{}^r}\Big(\delta_3 \underbrace{\sum_{s_i}, \cdots, \sum_{s_j}}_{\substack{r\text{-2 fold}\\ i<\cdots<j}} \bar{Y}^2_{\underbrace{s_i,\cdots,s_j}_{\substack{r\text{-2}\\ \text{subsc:ipts}\\ i<\cdots<j}}}\Big) - \cdots + (-1)^{r-1} \mathop{S}\limits_{1}^{r}\Big(\delta_r \sum_{s_i} \bar{Y}^2_{s_i}\Big)$$
$$+ (-1)^r N \bar{Y}^2$$

where $i, \cdots, j = 1, 2, \cdots, k$; k is the number of classifications in the whole study; r is the number of classifications under calculation; s_i (or s_j) $= 1, 2, \cdots, p_i$ (or p_j); δ_m is so determined that

$$(4.04)\quad \sum_{s_i}, \cdots, \sum_{s_j} \delta_m = N = p_1p_2, \cdots, p_kn \quad (m = 1, \cdots, r)$$

and throughout the general expression the summations and the subscripts which are not connected with the classifications under calculation should be ruled out. For instance, if we calculate the sum of squares for the interaction between A and B, the formula becomes

$$(4.05) \quad p_3, \cdots, p_k n \sum_{s_1} \sum_{s_2} \bar{Y}^2_{s_1 s_2}, \cdots, \\ - \left(p_2 p_3, \cdots, p_k n \sum_{s_1} \bar{Y}^2_{s_1}, \cdots, + p_1 p_3, \cdots, p_k n \sum_{s_2} \bar{Y}^2_{s_2}, \cdots\right) \\ + N\bar{Y}^2 \cdots$$

For another example, if we calculate the sum of squares for the main effect A, the formula becomes[4]

$$(4.06) \qquad p_2, \cdots, p_k n \sum_{s_1} \bar{Y}^2_{s_1}, \cdots, - N\bar{Y}^2, \cdots$$

References

1. Cochran, W. G., "Some Consequences When the Assumptions for the Analyses of Variance Are Not Satisfied," *Biometrics*, Vol. 3 (1947), pp. 22–38.
2. Eden, T. and Yates, F., "On the Validity of Fisher's Z-Test When Applied to an Actual Example of Non-Normal Data," *Journal of Agricultural Science*, Vol. XXIII (1933), pp. 6–17.
3. Eisenhart, Churchill, "The Assumptions Underlying the Analysis of Variance," *Biometrics*, Vol. 3 (1947), pp. 1–21.
4. Fisher, R. A., *The Design of Experiments*, 2d ed. London: Oliver and Boyd, 1937.
5. ———, "On a Distribution Yielding the Error Functions of Several Well-known Statistics," *Proceedings of the International Mathematical Congress*, Toronto, 1924, pp. 805–813.
6. ———, *Statistical Methods for Research Workers*, 10th ed. London: Oliver and Boyd, 1946.
7. ———, and Mackenzie, W. A., "Studies in Crop Variation. II: The Manurial Response of Different Potato Varieties," *Journal of Agricultural Science*, Vol. XIII (1923), pp. 311–320.
8. Pearson, E. S., *The Application of Statistical Methods to Industrial Standardization and Quality Control.* British Standards Institution No. 600, 1935.

[4] It is not possible to include either the mathematical solution of the problem of the analysis of covariance or the illustrations of specific problems prepared to show the use of the general expressions given in this section. The geometric representations of the analysis of variance have also been developed. The interested reader may secure mimeographed copies of these supplements by writing to the author.

CHAPTER XI

APPLICATIONS OF THE ANALYSIS OF VARIANCE AND COVARIANCE METHOD

We shall now apply the method of analysis of variance and covariance to a number of the simpler practical problems met with by the research worker. The application to some of the more complex types of situations in which these methods are indispensable will be made in the sequel to the results of specific experimental designs.

We shall proceed first by applying the principles presented in Chapter X to the mathematical solution of the problem. We shall then carry out the necessary calculations for the solution and interpretation. We begin with the simplest case of the analysis of variance, where there is a single criterion of classification.

Problem XI.1. Single classification with equal representation in classes. Let us take the problem of measuring the resemblance in intelligence of identical twins reared apart as reported by Newman, Freeman, and Holzinger (Ref. 7). The data in the form in which we shall use them in this analysis are given in Table 56. We must first see if we can translate our problem into mathematical language. If it is amenable to such a translation, it can be expressed mathematically as a problem of testing statistical hypotheses. Mathematically, the relationship may be expressed thus:

$$X_{it} = A + C_t + z_{it} \tag{11.01}$$

where $i = 1, 2$; $t = 1, 2, 3, \cdots, n(= 19)$; X_{it} is the mental age of the ith member of the tth pair of twins; A is a measure of the common mental age of the group tested; C_t is a measure of the mental age of the tth twin pair; z_{it} is the measure of the random effects. The restriction is

$$\sum_t C_t = 0 \tag{11.02}$$

We must first test the hypothesis that the variability of the mental-age scores is the same for all twin pairs, since this is the fundamental assumption underlying the analysis of variance. The hypothesis may be written

$$H_0{:}\sigma_t = \sigma \tag{11.03}$$

where σ_t denotes the standard deviation of the tth twin pair. This hypothesis is tested by means of the L-test (see page 82). The calcula-

TABLE 56
Mental Ages of 19 Pairs of Identical Twins Reared Apart

Twin pair (t)	Mental age		Sum $X_{1t} + X_{2t}$	Difference $\lvert X_{1t} - X_{2t}\rvert$
	X_{1t}	X_{2t}		
1	163	186	349	23
2	126	149	275	23
3	191	194	385	3
4	170	204	374	34
5	171	178	349	7
6	195	180	375	15
7	170	172	342	2
8	170	142	312	28
9	195	185	380	10
10	187	195	382	8
11	176	222	398	46
12	223	210	433	13
13	181	182	363	1
14	164	161	325	3
15	175	171	346	4
16	123	120	243	3
17	192	175	367	17
18	184	148	332	36
19	168	151	319	17
Sum	3,324	3,325	6,649	
Sum of squares	590,946	593,531	2,361,311	7,643

tions are carried out as indicated in Table 57. With a value of $L_1 = .117$, $k = 19$, and d.f. $= 1$, we refer to Nayer's table (Table V, Appendix) and find that our value is greater than the table value ($L_{1.01} = .096$). Hence, we may accept the hypothesis H_0 at the 1 per cent level.[1] We can now proceed to apply the analysis of variance method.

We use the maximum-likelihood procedure of estimating the sums of squares of the different components as shown below. We first write

$$\phi = \sum_i \sum_t (X_{it} - A - C_t)^2 + 2\lambda \sum_t C_t \tag{11.04}$$

where λ is the multiplier of Lagrange. Minimizing ϕ with respect to A, C_t, and λ, that is, differentiating partially with respect to A, C_t, and λ, equating the resulting equations to zero, and solving them for the values A, C_t, and λ, we obtain

[1] It should be pointed out that for the case $n = 2$, the L_1-test may sometimes be indecisive. We are accepting the hypothesis here at the 1 per cent level.

$$A = \frac{1}{2n}\sum_i\sum_t X_{it} = \bar{X}_{..} \qquad (11.05)$$

$$C_t = \tfrac{1}{2}\sum_i X_{it} - \bar{X}_{..} = \bar{X}_{.t} - \bar{X}_{..} \qquad (11.06)$$

$$\lambda = 0 \qquad (11.07)$$

TABLE 57

CALCULATION OF L_1 IN TESTING $H_0{:}\sigma_t = \sigma$

n_t	f_t	$\log n_t$	$n_t \log n_t$	$\theta_t' = \sum_i (X_{it} - \bar{X}_{.t})^2$	$\log \theta_t'$	$n_t \log \theta_t'$
2	1	.30103	.60206	264.5	2.42243	4.84486
2	1	.30103	.60206	264.5	2.42243	4.84486
2	1	.30103	.60206	4.5	.65321	1.30642
2	1	.30103	.60206	578.0	2.76193	5.52386
2	1	.30103	.60206	24.5	1.38917	2.77834
2	1	.30103	.60206	112.5	2.05115	4.10230
2	1	.30103	.60306	2.0	.30103	.60206
2	1	.30103	.60206	392.0	2.59329	5.18658
2	1	.30103	.60206	50.0	1.69897	3.39794
2	1	.30103	.60206	32.0	1.50515	3.01030
2	1	.30103	.60206	1058.0	3.02449	6.04898
2	1	.30103	.60206	84.5	1.92686	3.85372
2	1	.30103	.60206	.5	−.30103	−.60206
2	1	.30103	.60206	4.5	.65321	1.30642
2	1	.30103	.60206	8.0	.90309	1.80618
2	1	.30103	.60206	4.5	.65321	1.30642
2	1	.30103	.60206	144.5	2.15987	4.31974
2	1	.30103	.60206	648.0	2.81158	5.62316
2	1	.30103	.60206	144.5	2.15987	4.31974
$N = 38$		$\log N = 1.57978$	11.43914	3821.5	$\log \sum_t \theta_t' = 3.58224$	63.57982

Substituting these values in (11.04) to obtain the absolute minimum value of ϕ, we have

$$\chi_a^2 = \sum_i \sum_t (X_{it} - \bar{X}_{.t})^2 \qquad (11.08)$$

which is the basis for testing the following hypothesis:

$$H_1{:}E(C_t) = 0 \quad \left(\begin{array}{l} E \text{ is the notation for expectation} \\ \text{of a parameter} \end{array}\right) \qquad (11.09)$$

that is, the hypothesis that the mental age of an individual is independent of the particular twin pair to which the individual belongs. If the hypothesis is true, then (11.04) becomes

$$\phi = \sum_i \sum_t (X_{it} - A)^2 \qquad (11.10)$$

Minimizing with respect to A and substituting the obtained value, $A = \bar{X}_{..}$ in (11.10), we obtain the relative minimum:

$$\chi_r^2 = \sum_i \sum_t (X_{it} - \bar{X}_{..})^2 = \sum_i \sum_t (X_{it} - \bar{X}_{.t})^2 + \sum_i \sum_t (\bar{X}_{.t} - \bar{X}_{..})^2 \tag{11.11}$$

$$= \chi_a^2 + \chi_1^2 \tag{11.12}$$

where χ_a^2 is the estimate of sum of squares for "within" and χ_1^2 is the estimate of the sum of squares for "between." Then the test of H_1 is given by

$$F = \frac{n}{n-1}\frac{\chi_1^2}{\chi_a^2} \tag{11.13}$$

with $n_1 = n - 1$ and $n_2 = n$. For purposes of calculation it is simpler to write χ_a^2 and χ_1^2 in the form

$$\chi_a^2 = \tfrac{1}{2} \sum_t (X_{1t} - X_{2t})^2 \tag{11.14}$$

$$\chi_1^2 = \frac{1}{2}\left[\sum_t (X_{1t} + X_{2t})^2 - \frac{\left(\sum_i \sum_t X_{it}\right)^2}{n}\right] \tag{11.15}$$

Calculations may be checked by

$$\chi_r^2 = \sum_i \sum_t X_{it}^2 - \frac{\left(\sum_i \sum_t X_{it}\right)^2}{2n} \tag{11.16}$$

Separately, and using the identity,

$$\chi_r^2 = \chi_a^2 + \chi_1^2 \tag{11.17}$$

The efficient way to calculate the necessary values is shown in Table 56; we first form the sum and difference for each pair of values. We then calculate the sum and sum of squares for each column, except the last, where the sum of squares only is needed. By this method we secure a check on the calculations at each stage.

From the last two rows of Table 56, we obtain

$$\sum_t (X_{1t} - X_{2t})^2 = 7643 \tag{11.18}$$

$$\sum_t (X_{1t} + X_{2t})^2 = 2{,}361{,}311 \tag{11.19}$$

$$\sum_i \sum_t X_{it} = 6649 \tag{11.20}$$

$$\sum_i \sum_t X_{it}^2 = \tfrac{1}{2}\left[\sum_t (X_{1t} + X_{2t})^2 + \sum_t (X_{1t} - X_{2t})^2\right] = 1{,}184{,}477 \tag{11.21}$$

Substituting these values in (11.14), (11.15), and (11.16), we have

$$\begin{aligned} \chi_a^2 &= 3\,821.5 \\ \chi_1^2 &= 17{,}255.5 \\ \chi_r^2 &= 21{,}077.0 \end{aligned}$$

We now place all of these values in one table as shown in Table 58.

TABLE 58
ANALYSIS OF VARIANCE OF MENTAL AGES OF IDENTICAL TWINS REARED APART

Source of variation	D.F.	Sum of squares	Mean square	F	Hypothesis
Within pairs	19	3,821.5	201.132	....	....
Between pairs	18	17,255.5	958.638	4.766	Rej.
Total	37	21,077.0			

Referring to the F-table with $n_1 = 18$ and $n_2 = 19$, we find that the obtained value of F is significant at the 1 per cent level. This statement means that the mental age of an individual is not independent of the twin pair to which he belongs, or that there is a significant difference among the means of the 19 twin pairs. Another interpretation is that the intraclass correlation between twins is significantly greater than zero. Intraclass correlation is discussed below.

Fisher (Reference 3) has shown that an unbiased estimate of the intraclass correlation, r', can be obtained from the relation

$$F = \frac{1 + (k - 1)r'}{1 - r'} \tag{11.22}$$

where k is the number in a group or class. Where $k = 2$,

$$F = \frac{1 + r'}{1 - r'} \tag{11.23}$$

Thus in our problem, $\dfrac{1 + r'}{1 - r'} = \dfrac{958.638}{201.132} = 4.7662$

$$r' = .653$$

Also, $$r' = \frac{958.638 - 201.132}{958.638 + (2 - 1)(201.132)} = .653$$

When there are equal numbers in the classes or groups, the variation of the class means relative to the variation of the individuals within the classes is measured by the intraclass correlation. If the class means differ significantly, a significant positive intraclass correlation is indicated; when the mean square between classes equals that within classes, the correlation is zero; and if the mean square between classes is less than that within classes, the intraclass correlation is negative.

Problem XI.2. Testing the homogeneity of multiple groups of measurements. We shall apply the analysis-of-variance method to test the homogeneity of 6 sections in college zoology with respect to their achievement as measured by a final examination. The basic data are given in Table 59.

Denote by X_{st} the score of the tth student in the sth section. The basic assumption in the analysis is that we may write

$$X_{st} = A + B_s + z_{st} \qquad (11.24)$$

where $s = 1, 2, \cdots, \text{k}$; $t = 1, 2, \cdots, n_s$; n_s denotes the number of students in the sth section, and k denotes the number of sections. A is a measure of the achievement of all the students and is defined as the mean score for all individuals and sections; B_s is a measure of the achievement of the sth section; z_{st} is a measure of random effects, assumed to be normally distributed about zero with constant standard deviation, σ. The restriction is

$$\sum_s B_s = 0 \qquad (11.25)$$

In assuming that σ is constant, we are assuming that the variability of the scores is the same for each section. This assumption may not be fulfilled in practice, and hence, we must first test the hypothesis

$$H_0 : \sigma_s = \sigma \qquad (11.26)$$

TABLE 59

SUMS AND SUMS OF SQUARES OF SCORES FOR EACH SECTION IN COLLEGE ZOOLOGY

Section	No. of students n_s	Sum of scores ΣX	Sum of squares of scores ΣX^2	Sum of squares about means $\Sigma X^2 - \frac{(\Sigma X)^2}{n_s}$
I	145	23,025	3,759,061	102,849.7931
II	91	13,529	2,065,833	54,472.1099
III	84	13,127	2,130,435	79,028.7024
IV	127	18,825	2,912,131	121,732.3779
V	46	6,828	1,071,968	58,455.3043
VI	82	12,889	2,108,159	82,228.2560
Total	575	88,223	14,047,587 →	511,417.0036

where σ_s denotes the standard deviation of the scores in the sth section. If this hypothesis is accepted, we conclude that there is no difference in variability among the sections and then proceed to test the other hypothesis. If we reject the hypothesis H_0, we cannot make an exact test of another hypothesis.

The test of the hypothesis H_0 may be made as follows. We calculate

$$L_1 = \prod_s \left(\frac{N}{n_s}\right)^{\frac{n_s}{N}} \prod_s \left(\frac{\theta'_s}{\sum_s \theta'_s}\right)^{\frac{n_s}{N}} \qquad (11.27)$$

where $N = \sum_s n_s$, Π denotes the product, and θ'_s denotes the "within" sections sum of squares for the sth section. We refer to Nayer's tables of the L_1-distribution with $k = 6$ and d.f. equal to the harmonic mean of f_s, where f_s denotes the d.f. associated with θ'_s in the sth section. The rule to be followed in using these tables is to reject the hypothesis when the calculated value of L_1 is less than the corresponding 1 per cent point given in the table.

The computation of the L_1 for the 6 sections is carried out as shown in Table 60.

TABLE 60

CALCULATION OF L_1 FOR THE TEST OF THE HYPOTHESIS $H_0: \sigma_s = \sigma$

n_s	$\log n_s$	$n_s \log n_s$	θ_s'	$\log \theta_s'$	$n_s \log \theta_s'$
145	2.16137	313.39865	102,849.7931	5.01221	726.77045
91	1.95904	178.27264	54,472.1099	4.73618	430.99238
84	1.92428	161.63952	79,028.7024	4.89778	411.41352
127	2.10380	267.18260	121,732.3779	5.08541	645.84707
46	1.66276	76.48696	58,455.3043	4.76682	219.27372
82	1.91381	156.93242	82,228.2560	4.91502	403.03164
$N = 575$	$\log N = 2.75969$	1153.91279	$\Sigma\theta_s' = 498{,}766.5436$	$\log \Sigma\theta_s' = 5.69790$	2837.32878

To find the value of L_1, we calculate the value of $\log L_1$, where

$$\log L_1 = \log N - \frac{1}{N}\sum_t n_s \log n_s + \frac{1}{N}\sum_s n_s \log \theta'_s - \log\left(\sum_s \theta'_s\right) \text{ and then}$$

find L_1 from a table of antilogarithms.

$$\begin{aligned} \text{Here,}\quad \log L_1 &= 2.75967 - 2.00680 + 4.93448 - 5.69790 \\ &= 9.98945 - 10 \\ L_1 &= .9760 \end{aligned}$$

$$\text{The harmonic mean of } f_s = \frac{6}{\frac{1}{144} + \frac{1}{90} + \frac{1}{83} + \frac{1}{126} + \frac{1}{45} + \frac{1}{81}} = 82.64$$

Referring to Nayer's tables with $k = 6$ and d.f. $= 83$, we find that $P > .05$. We accept the hypothesis H_0 and conclude that the sections are of equal variability. Consequently, we can proceed to the analysis of variance.

The next step is to estimate the sum of squares for "within." By the method of maximum likelihood, we obtain

$$\phi = \sum_s \sum_t (X_{st} - A - B_s)^2 + 2\lambda \sum_s B_s \qquad (11.28)$$

Differentiating ϕ partially with respect to A, B_s, and λ, equating these equations to zero, and solving the resulting equations for the values of A, B_s, and λ, we obtain

$$A = \frac{1}{N}\sum_s \sum_t X_{st} = \bar{X}.. \tag{11.29}$$

$$B_s = \frac{1}{n_s}\sum_t X_{st} - A = \bar{X}_s.. - \bar{X}.. \tag{11.30}$$

$$\lambda = 0 \tag{11.31}$$

Substituting these values in (11.28) to obtain the absolute minimum value of ϕ, we have

$$\chi_a^2 = \sum_b \sum_t X_{st}^2 - \sum_s \left[\frac{\left(\sum_t X_{st}\right)^2}{n_s}\right] \tag{11.32}$$

which is the basis of testing the following hypothesis:

$$H_1 : E(B_s) = 0 \quad \left(\begin{array}{l} E \text{ is the notation for expectation of} \\ \text{a parameter} \end{array}\right) \tag{11.33}$$

that is, the hypothesis that the sections are equal in achievement. If the hypothesis is true, then (11.28) becomes

$$\phi = \sum_s \sum_t (X_{st} - A)^2 \tag{11.34}$$

Minimizing with respect to A and substituting the obtained value of $A = \bar{X}..$ in (11.34), we obtain the relative minimum:

$$\chi_r^2 = \sum_s \sum_t X_{st}^2 - \frac{\left(\sum_s \sum_t X_{st}\right)^2}{N} = \chi_a^2 + \sum_s \frac{\left(\sum_t X_{st}\right)^2}{n_s} - \frac{\left(\sum_s \sum_t X_{st}\right)^2}{N}\Bigg]$$
$$= \chi_a^2 + \chi_1^2 \tag{11.35}$$

where χ_a^2 is the estimate of sum of squares for "within" and χ_1^2 is the estimate of sum of squares for "between." Then the test of H_1 is given by

$$F = \frac{N-6}{5}\frac{\chi_1^2}{\chi_a^2} \tag{11.36}$$

with $n_1 = 5$ and $n_2 = N - 6$.

The "within" sum of squares may be obtained directly from the last row of Table 60, $\sum_s \theta_s' = 498{,}766.5436$. The "between" sum of squares is calculated from the totals given in the third column of Table 59 as follows:

$$\frac{(23{,}025)^2}{145} + \frac{(13{,}529)^2}{91} + \frac{(13{,}127)^2}{84} + \frac{(18{,}825)^2}{127} + \frac{(6828)^2}{46} + \frac{(12{,}889)^2}{82} - \frac{(88{,}223)^2}{575} = 12{,}650.4599$$

The total sum of squares is

$$\sum_s \sum_t X_{st}^2 - \frac{\left(\sum_s \sum_t X_{st}\right)^2}{N} = 14{,}047{,}587 - \frac{(88{,}223)^2}{575} = 511{,}417.0036$$

To test the hypothesis H_1, we calculate

$$F = \frac{569}{5} \cdot \frac{12{,}650.46}{498{,}766.5436} = \frac{2530.092}{876.567} = 2.886$$

Referring to the F tables with $n_1 = 5$ and $n_2 = 569$, we find that $.05 > P > .01$. Statistically, the acceptance of H_1 remains in doubt. We may state that the differences among the means of sections are significant at the 5 per cent level but not significant at the 1 per cent level. The results are summarized in Table 61.

TABLE 61

ANALYSIS OF VARIANCE OF SCORES IN DIFFERENT SECTIONS IN ZOOLOGY

Variance	D.F.	Sum of squares	Mean square	F	Hypothesis
Between sections	5	12,650.46	2530.092	2.886	Remains in doubt
Within sections	569	498,766.5436	876.567		
Total	574	511,417.0036			

When a significant difference has been found among the means of the sections, it may be of interest to make special comparisons between the means of any two of the sections. Here the usual test of significance between means cannot be applied, since the two means cannot now be regarded as randomly drawn from a normal population. Fisher (Ref. 2) has suggested a test taking into account the probability of a random sampling based on binomial theory. This method will be illustrated by comparing the mean of the highest with that of the lowest section. The analysis of variance could be used, but we shall apply the t-test with the modifications indicated.

We shall test the significance of the difference between the means of Section I and of Section IV, 158.8 and 148.2, respectively. We find that

$$t_0 = \frac{(\bar{X}_1 - \bar{X}_2)}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{158.8 - 148.2}{28.84\sqrt{\frac{1}{145} + \frac{1}{127}}} = 3.02$$

for $n = 270$ the corresponding $P < .01$. The difference selected, however, is only 1 of 15 that might be found among the means of the six sections. The required probability for the selected difference to be significant is set, therefore, not as 1 in 100 but as 1 in (15)(100) = 1500. Since the probability corresponding to the observed value of t is about .0024, or 2.4 in 1000, it is greater than 1.5 in 1000 and therefore is regarded as not significant.

Problem XI.3. An application of the analysis of covariance. The process of applying the analysis of covariance consists in breaking up the sum of products into parts assignable to different factors. This is comparable to the process of breaking up the sum of squares in the case of the analysis of variance.

We shall apply the method of the analysis of variance and covariance to the combined analysis of mental-age scores and educational-age scores, as measured by the Stanford Achievement Test, in the case of the 19 pairs of identical twins reared apart.

Let X_{it} denote the educational age of the ith member of the tth twin pair and Y_{it} the mental age of the ith member of the tth twin pair. We may then write

$$X_{it} = A + C_t + z_{it} \tag{11.37}$$

and

$$Y_{it} = B + D_t + z'_{it} \tag{11.38}$$

with restrictions

$$\sum_t C_t = 0 \tag{11.39}$$

$$\sum_t D_t = 0 \tag{11.40}$$

where $i = 1, 2$; $t = 1, 2, \cdots, n$; where n is the number of twin pairs.

The difference between the educational ages of the pairs of identical twins may be due partly or wholly to differences in mental age. The problem is to find out what part of these differences may be assigned to differences in mental age and to adjust the analysis accordingly. We wish to find out whether there is a difference in achievement of the identical twins when they may be regarded as of the same mental age.

If we may assume that there is a linear relationship between educational age and mental age,[2] we may write, since Y_{it} denotes the mental age of the ith member of the tth twin pair,

$$X_{it} = a + bY_{it} + S_{it} \tag{11.41}$$

where a and b are parameters to be estimated from the data; b is the regression coefficient of educational age on mental age; S_{it} is the measure of the differences between the educational ages of members of the same

[2] For the test of linearity see page 240.

twin pairs and of those between the educational ages of pairs of twins not attributable to the factor of mental age.

As formerly, we minimize:

$$\phi = \sum_i \sum_t (X_{it} - a - bY_{it})^2 + 2\lambda_1 \sum_t C_t + 2\lambda_2 \sum_t D_t \qquad (11.42)$$

with regard to a, b, λ_1, and λ_2 to obtain the relative minimum of ϕ, χ_r^2.

Solving for a, b, λ_1, and λ_2, we have

$$a = \frac{1}{2n}\left(\sum_i \sum_t X_{it} - b \sum_i \sum_t Y_{it}\right) \qquad (11.43)$$

$$b = \frac{\sum_t (X_{1t}Y_{1t} + X_{2t}Y_{2t}) - \frac{\left[\sum_t (X_{1t} + X_{2t})\right]\left[\sum_t (Y_{1t} + Y_{2t})\right]}{2n}}{\sum_i \sum_t Y_{it}^2 - \frac{\left(\sum_i \sum_t Y_{it}\right)^2}{2n}} \qquad (11.44)$$

$$\lambda_1 = 0 \qquad (11.45)$$

$$\lambda_2 = 0 \qquad (11.46)$$

Substituting the values of a, b, λ_1, and λ_2 into Equation (11.42), we obtain

$$\chi_r^2 = \frac{1}{2}\sum_t (X_{1t} - X_{2t})^2 + \frac{1}{2}\left[\sum_t (X_{1t} + X_{2t})^2 - \frac{\left(\sum_i \sum_t X_{it}\right)^2}{n}\right] - \frac{\left\{\sum_t (X_{1t}Y_{1t} + X_{2t}Y_{2t}) - \frac{\left[\sum_t (X_{1t} + X_{2t})\right]\left[\sum_t (Y_{1t} + Y_{2t})\right]}{2n}\right\}}{\sum_i \sum_t Y_{it}^2 - \frac{\left(\sum_i \sum_t Y_{it}\right)^2}{2n}} \qquad (11.47)$$

$$= \chi_a^2 + \chi_1^2, \text{ say} \qquad (11.48)$$

The proportion of the variance attributable to mental age is

$$l = \frac{\left\{\sum_t (X_{1t}Y_{1t} + X_{2t}Y_{2t}) - \frac{\left[\sum_t (X_{1t} + X_{2t})\right]\left[\sum_t (Y_{1t} + Y_{2})\right]}{2n}\right\}^2}{\sum_i \sum_t Y_{it}^2 - \frac{\left(\sum_i \sum_t Y\right)^2}{2n}} \qquad (11.49)$$

To obtain χ_1^2, we subtract l from the "between" pairs sum of squares, since the other two quantities in (11.47) are the "within" and "between" pairs sums of squares for educational age.

The necessary calculations may be efficiently carried out if the data are arranged as shown in Table 62.

The results are presented in tabular form in Table 63.

TABLE 62
EDUCATIONAL AND MENTAL AGES OF 19 PAIRS OF IDENTICAL TWINS REARED APART

Twin pair	Educational age X_{it}		$\lvert X_{1t} - X_{2t}\rvert$	$X_{1t} + X_{2t}$	Mental age Y_{it}		$X_{1t} \cdot Y_{1t}$	$X_{2t} \cdot Y_{2t}$	$(X_{1t} + X_{2t})$ $(Y_{1t} + Y_{2t})$
	X_{1t}	X_{2t}			Y_{1t}	Y_{2t}			
1	181	200	19	381	163	186	29,503	37,200	132,969
2	131	169	38	300	126	149	16,506	25,181	82,500
3	205	189	16	394	191	194	39,155	36,666	151,690
4	173	207	34	380	170	204	29,410	42,228	142,120
5	176	182	6	358	171	178	30,096	32,396	124,942
6	151	155	4	306	195	180	29,445	27,900	114,750
7	191	189	2	380	170	172	32,470	32,508	129,960
8	175	162	13	337	170	142	29,750	23,004	105,144
9	210	202	8	412	195	185	40,950	37,370	156,560
10	181	200	19	381	187	195	33,847	39,000	145,542
11	157	226	69	383	176	222	27,632	50,172	152,434
12	224	210	14	434	223	210	49,952	44,100	187,922
13	196	189	7	385	181	182	35,476	34,398	139,755
14	176	159	17	335	164	161	28,864	25,599	108,875
15	159	161	2	320	175	171	27,825	27,531	110,720
16	130	131	1	261	123	120	15,990	15,720	63,423
17	176	176	0	352	192	175	33,792	30,800	129,184
18	192	157	35	349	184	148	35,328	23,236	115,868
19	177	172	5	349	168	151	29,736	25,972	111,331
Total	3,361	3,436		6,797	3,324	3,325	595,727	610,981	2,405,689
Sum of squares	605,187	631,518	10,417	2,462,993	590,946	593,531	1,206,708		

TABLE 63
ANALYSIS OF VARIANCE AND COVARIANCE OF EDUCATIONAL AND MENTAL AGES

Variance	D.F.	Sums of squares		Sums of Products M.A. times E.A.	Regression coefficient	Correlation coefficient
		Mental age	Educational age			
Between pairs of twins	18	17,255.5	15,727.8	13,548.369	0.785	.822
Within pairs of twins	19	3,821.5	5,208.5	3,863.500	1.011	.866
Total	37	21,077.0	20,936.3	17,411.900	0.827	.829

The quantities entered in Table 63 are calculated as follows.

Educational age:

$$\text{Between pairs: } \tfrac{1}{2}\sum_t (X_{1t} + X_{2t})^2 - \tfrac{1}{38}\Big(\sum_i \sum_t X_{it}\Big)^2 = 15{,}727.8$$

$$\text{Within pairs: } \tfrac{1}{2}\sum_t (X_{1t} - X_{2t})^2 = 5208.5$$

$$\text{Total: } \sum_i \sum_t X_{it}^2 - \tfrac{1}{38}\Big(\sum_i \sum_t X_{it}\Big)^2 = 20{,}936.3$$

Mental age:

$$\text{Between pairs: } \tfrac{1}{2}\sum_t (Y_{1t} + Y_{2t})^2 - \tfrac{1}{38}\Big(\sum_i \sum_t Y_{it}\Big)^2 = 17{,}255.5$$

$$\text{Within pairs: } \tfrac{1}{2}\sum_t (Y_{1t} - Y_{2t})^2 = 3821.5$$

$$\text{Total: } \sum_i \sum_t Y_{it}^2 - \tfrac{1}{38}\Big(\sum_i \sum_t Y_{it}\Big)^2 = 21{,}077.0$$

Products of educational age by mental age:

$$\begin{aligned}\text{Between pairs: } & \tfrac{1}{2}\sum_t (X_{1t} + X_{2t})(Y_{1t} + Y_{2t}) \\ & - \tfrac{1}{38}\Big[\sum_t (X_{1t} + X_{2t})\Big]\Big[\sum_t (Y_{1t} + Y_{2t})\Big] \\ & = \tfrac{1}{2}(2{,}405{,}689) - \tfrac{1}{38}[6797][6649] = 13{,}548.4\end{aligned}$$

$$\begin{aligned}\text{Within pairs: } & \sum_t X_{1t}Y_{1t} + \sum_t X_{2t}Y_{2t} - \tfrac{1}{2}\sum_t (X_{1t} + X_{2t})(Y_{1t} + Y_{2t}) \\ & = 1{,}206{,}708 - \tfrac{1}{2}(2{,}405{,}689) = 3{,}863.5\end{aligned}$$

$$\begin{aligned}\text{Total: } & \sum_t X_{1t}Y_{1t} + \sum_t X_{2t}Y_{2t} \\ & - \tfrac{1}{38}\Big[\sum_t (X_{1t} + X_{2t})\Big]\Big[\sum_t (Y_{1t} + Y_{2t})\Big] \\ & = 1{,}206{,}708 - \tfrac{1}{38}[6797][6649] = 17{,}411.9\end{aligned}$$

Two methods for adjusting the sum of squares of educational ages are given. The first method makes possible a more nearly exact test of significance. The adjusted sums of squares are obtained by adjusting the "within" pairs and "total" each with its own regression coefficient and subtracting to find the adjusted "between pairs" sum of squares. This method is depicted in Table 64. The adjustments are as follows:

$$\text{Within pairs: } 5208.5 - \frac{(3863.5)^2}{3821.5} = 1302.5$$

$$\text{Total: } 20{,}936.3 - \frac{(17{,}411.9)^2}{21{,}077} = 6552.5$$

$$\text{Between pairs: } 6552.5 - 1302.5 = 5249.7$$

TABLE 64

ANALYSIS OF VARIANCE OF EDUCATIONAL AGE SCORES OF THE 19 PAIRS OF IDENTICAL TWINS—METHOD 1 ORIGINAL AND ADJUSTED SUMS OF SQUARES AND MEAN SQUARES

Variance	D.F.	Original analysis		D.F.	Adjusted analysis	
		Sum of squares	Mean square		Sum of squares	Mean square
Between pairs	18	15,727.8	873.767	18	5249.7	291.65
Within pairs	19	5,208.5	274.132	18	1302.5	72.36
Total	37	20,936.3	565.846	36	6552.2	

A second method of adjusting the sum of squares is shown in Table 65. Both the "between pairs" and the "within pairs" sums of squares are adjusted by the use of the "within pairs" regression coefficient:

$$\begin{aligned}\Sigma(x - by)^2 &= \Sigma x^2 - 2b\Sigma xy + b^2\Sigma y^2 \\ &= 15{,}727.8 - 2\frac{3863.5}{3821.5}(13{,}548.4) + \left(\frac{3863.5}{3821.5}\right)^2 (17{,}255.5) \\ &= 15{,}727.8 - 2.02198(13{,}548.4) + 1.02210(17{,}255.5) \\ &= 15{,}727.8 - 27{,}394.5938 - 17{,}636.8466 = 5970.053\end{aligned}$$

In certain cases it may be necessary to adjust each sum of squares with its own regression coefficient (Ref. 5).

TABLE 65

ANALYSIS OF VARIANCE OF EDUCATIONAL AGE SCORES OF THE 19 PAIRS OF IDENTICAL TWINS—METHOD 2 ORIGINAL AND ADJUSTED SUMS OF SQUARES AND MEAN SQUARES

Variance	D.F.	Original analysis		D.F.	Adjusted analysis	
		Sum of squares	Mean square		Sum of squares	Mean square
Between pairs	18	15,727.8	873.767	18	5970.053	331.670
Within pairs	19	5,208.5	274.132	18	1302.500	72.361
Total	37	20,936.3	565.846	36		

The adjusted between pairs sum of squares and mean square give a measure of the difference between twin pairs in educational age freed from the influence of mental age. To test the hypothesis that these adjusted differences are zero, we calculate:

$$z_0 = \frac{1}{2}\log_e \frac{\text{mean square between pairs}}{\text{mean square within pairs}}$$

and refer to Fisher's tables of z with degrees of freedom $n_1 = n - 1$ and $n_2 = n - 1$, where n is the number of twin pairs. In our example we have

$$z_0 = \frac{1}{2}\log_e \frac{291.65}{72.36} \text{ or } \frac{1}{2}\log_e 4.03 = .697 \qquad \text{(Table 64)}$$

or

$$z_0 = \frac{1}{2}\log_e \frac{331.670}{72.36} \text{ or } \frac{1}{2}\log_e 4.584 = .761 \qquad \text{(Table 65)}$$

From Fisher's tables of z, entered with degrees of freedom $n_1 = 18$ and $n_2 = 18$, we find that z_0 is greater than the value given in the table at the 1 per cent level. We could also have used the tables of Snedecor's F. We reject the hypothesis and conclude that when the factor of mental age is removed, the means of the educational ages of twin pairs differ significantly.

We obtain three measures of the degree of relationship between educational age and mental age from the results of Table 63. From the first row, for between pairs, we have

$$r_1 = \frac{13{,}548.369}{\sqrt{(17{,}255.5)(15{,}727.8)}} = \frac{13{,}548.369}{\sqrt{271{,}391{,}052.9}} = .822$$

From the second row, for within pairs, we have

$$r_2 = \frac{3863.5}{\sqrt{(3821.5)(5208.5)}} = .866$$

From the third row, for the total, we have

$$r_3 = \frac{17{,}411.9}{\sqrt{(21{,}077)(20{,}936.3)}} = \frac{17{,}411.9}{21{,}006} = .829$$

The second, $r_2 = .866$, is the best measure of the degree of relationship; in the third, $r_3 = .829$ the relationship is masked by the inclusion of the between-pairs differences in educational age and in mental age.

Problem XI.4. Test of the linearity of regression. The statistical study of the relationship between two or more variables involves consideration of the kind of relationship existing among them. Regression may be linear or nonlinear, and it is essential in any problem involving the use of regression to determine which particular kind best represents the observational data. The statistical method of correlation, particularly the product-moment correlation coefficient, involves the assumption of linearity of regression. The analysis of variance provides a straightforward method of testing the type of regression. Since linear regression is the type most often encountered, we shall consider here the problem of testing the linearity of regression (Ref. 5).[3]

[3] For other cases of polynomial equations and especially for the separation of sums of squares corresponding to individual degrees of freedom where the independent effects are represented by polynomials of different degree, see page 309.

We may take as a practical problem the case presented in Table 66, that of determining whether the relationship between the scores of the same individuals on two tests, one administered prior and the other subsequent to instruction, was linear in form. We shall also test another assumption underlying the product-moment correlation method, the homoscedasticity of variances of the different arrays, that is, if the variances of the different arrays are equal.

TABLE 66

CORRELATION TABLE FOR THE INITIAL AND FINAL SCORES OF 263 STUDENTS ON A TEST IN COLLEGE BIOLOGY

X (initial score)

		10 –	12 –	14 –	16 –	18 –	20 –	22 –	24 –	26 –	28 –	30 –	*F*
Y	44 –									1			1
	42 –			1									1
	40 –										1	1	2
F	38 –						2		1				3
i	36 –	1	1	1	2	4	3	4	4	4	5	3	32
n	34 –			2	2	5	6	3	6	5	3	2	34
a	32 –	1	2	1	2	5	5	3	4	8	4	1	36
l	30 –	1		1	4	3	6	5	5	8	4		37
	28 –			2	7	8	1	5	3	2	2	1	31
S	26 –	3	1	1	4	5	5	3	2	3	2	1	30
c	24 –	3	2	2	5	6	3	3	3	1	1		29
o	22 –	2	3	1	2	4	1	1					14
r	20 –	1	1	2	1	1	1					1	8
e	18 –				1			1	1				3
	16 –							1					1
	14 –								1				1
F		12	10	14	30	41	33	29	30	32	22	10	263

Let *X* and *Y* represent the scores on the initial and final tests, respectively. Then the regression function, when linear, is given by

$$\hat{Y} = a + b(X - \bar{X}) \tag{11.50}$$

where *a* and *b* are two parameter values, the value chosen for *a* being the mean, $\bar{Y}$, of the observed values *Y*, and the value given to *b* being the estimate of the regression coefficient of *Y* and *X*. $\hat{Y}$ is the expected value of *Y* for each *X*, and $\bar{X}$ is the mean of the *X* values.

In Table 66 the data are grouped, and we shall take as the selected values of *X* the mid-points of the several class intervals as shown in Table 67. It is observed in Table 66 that for each *X* the several values of *Y* form an array. Then, letting Y_{st} represent the score on the final test of the *t*th individual in the *s*th array, we have

$$\hat{Y}_{st} = A + B_s + z_{st} \tag{11.51}$$

where $t = 1, 2, \cdots, n_s$; $s = 1, 2, \cdots, k$; k denotes the number of arrays and n_s the number of individuals in the sth array. A is the mean of the scores of all individuals on the final test; B_s gives the measure of achievement on the final test of all individuals in the sth array; and z_{st} represents the measure of residual variation or the portion of Y_{st} attributable to random factors, such as errors of measurement, which are independent of X. z_{st} is assumed to be normally distributed about 0 with standard deviation σ, supposed to be the same for all arrays. The latter assumption will be tested first, by using the L_1-test. The sums and sums of squares of the scores on the final test in each array are given in Table 67. From Welch's formula for the L_1-test,

$$L_1 = \prod_s \left(\frac{N}{n_s}\right)^{\frac{n_s}{N}} \prod_s \left(\frac{\theta'_s}{\sum_s \theta'_s}\right)^{\frac{n_s}{N}} \tag{11.52}$$

we have

$$\log L_1 = \log N - \frac{1}{N} \sum_s n_s \log n_s + \frac{1}{N} \sum_s n_s \log \theta'_s - \log \left(\sum_s \theta'_s\right) \tag{11.53}$$

TABLE 67
SUM AND SUM OF SQUARES OF FINAL SCORES IN EACH ARRAY

Array No.	Value of X_s	n_s	Sum of scores $\sum_t Y_{st}$	Sum of squares of scores $\sum_t Y^2_{st}$	$\frac{\left(\sum_t Y_{st}\right)^2}{n_s}$	$\sum_t Y^2_{st} - \frac{\left(\sum_t Y_{st}\right)^2}{n_s}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	10.5	12	318.0	8,659.00	8,427.00	232.00
2	12.5	10	265.0	7,286.50	7,022.50	264.00
3	14.5	14	407.0	12,379.50	11,832.07	547.43
4	16.5	30	841.0	24,149.50	23,576.03	573.47
5	18.5	41	1190.5	35,408.25	34,565.62	842.63
6	20.5	33	1016.5	32,016.25	31,312.80	703.45
7	22.5	29	852.5	25,809.25	25,060.56	748.69
8	24.5	30	919.0	29,023.25	28,152.03	870.22
9	26.5	32	1028.0	33,484.00	33,024.50	459.50
10	28.5	22	713.0	23,447.50	23,107.68	339.82
11	30.5	10	325.0	10,930.50	10,562.50	368.00
Total		263	7875.5	242,593.50	236,643.29	5950.21

In our problem, as shown in Table 68, we find

$$\log L_1 = 2.4200 - 1.4219 + 2.7593 - 3.7745 = 9.9829 - 10$$

So we obtain $L_1 = .961$. Referring to Nayer's tables with $k = 11$ and harmonic mean,

$$f_s = \frac{11}{\frac{1}{11} + \frac{1}{9} + \frac{1}{13} + \frac{1}{29} + \frac{1}{40} + \frac{1}{32} + \frac{1}{28} + \frac{1}{29} + \frac{1}{31} + \frac{1}{21} + \frac{1}{9}}$$
$$= 18$$

TABLE 68
CALCULATION OF L_1 FOR THE TEST OF THE HYPOTHESIS $H_0{:}\sigma_s = \sigma$

f_s	n_s	$\log n_s$	$n_s \log n_s$	θ_s'	$\log \theta_s'$	$n_s \log \theta_s'$
11	12	1.0792	...	232.00	2.3655	...
9	10	1.0000	...	264.00	2.4216	...
13	14	1.1461	...	547.43	2.7383	...
29	30	1.4771	...	573.47	2.7585	...
40	41	1.6128	...	842.63	2.9256	...
32	33	1.5185	...	703.35	2.8472	...
28	29	1.4624	...	748.69	2.8742	...
29	30	1.4771	...	871.22	2.9401	...
31	32	1.5051	...	459.50	2.6623	...
21	22	1.3424	...	339.82	2.5313	...
9	10	1.0000	...	368.00	2.5658	...
	263	$\sum_s n_s \log n_s = 373.9627$		5950.21	$\sum_s n_s \log \theta_s' = 725.6954$	

we find that the value of L_1 is greater than the tabled value at the 5 per cent level, so we may assume that σ_s is constant. The first analysis of the scores for the final test is given in Table 69.

TABLE 69
ANALYSIS OF VARIANCE OF SCORES ON FINAL TEST

Source of variation	D.F.	Sum of squares	Mean square
Between means of arrays	10	812.49	81.249
Within arrays	252	5950.21	16.904
Total	262	6762.70	......

The analysis consists in breaking up the total sum of squares into two components. One component gives the mean-square estimate of the population variance between means of arrays and the other the mean-square estimate within arrays. The respective mean squares are given in the analysis-of-variance table. The sums of squares for each source of variation are obtained as follows, making use of the totals recorded in Table 67.

The within-rays sum of squares is the total of column (7), 5950.21; the between-means of arrays sum of squares is obtained from the totals of columns (4) and (6):

$$236{,}643.29 - \frac{(7875.5)^2}{263} = 812.49$$

and the total sum of squares is calculated from the totals of columns (4) and (5):

$$242{,}593.50 - \frac{(7875.5)^2}{263} = 6762.70$$

The hypothesis H_1, that the regression of Y on X is linear, is stated as follows:

$$H_1 : \hat{Y}_s = a + b(X_s - \bar{X}) \tag{11.54}$$

where $\hat{Y}_s$ is the expected value of Y for X_s, the sth value of X. If H_1· is accepted, then Equation (11.51) may be written

$$\hat{Y}_{st} = a + b(X_s - \bar{X}) \tag{11.55}$$

H_1 may then be tested in the conventional manner of testing a linear hypothesis. We have

$$\chi^2 = \sum_s \sum_t (Y_{st} - A - B_s)^2 \tag{11.56}$$

We then minimize χ^2 with respect to all parameters to get the absolute minimum χ^2_a. Thus:

$$\chi^2_a = \sum_s \sum_t Y^2_{st} - \sum_s \left[\frac{\left(\sum_t Y_{st}\right)^2}{n_s} \right] \tag{11.57}$$

which gives the sum of squares within arrays.

We then minimize χ^2 with respect to the parameters remaining under the assumption that H_1 is true. Thus, minimize

$$\chi^2 = \sum_s \sum_t [Y_{st} - a - b(X_s - \bar{X})]^2 \tag{11.58}$$

with respect to a and b to get the relative minimum, χ^2_r. We get

$$a = \frac{1}{\sum_s n_s} \sum_s \sum_t Y_{st} = \bar{Y}.. \tag{11.59}$$

$$b = \frac{\sum_s \left[(X_s - \bar{X}) \left(\sum_t Y_{st} \right) \right]}{\sum_s [n_s(X_s - \bar{X})^2]} \tag{11.60}$$

Then we may write

$$\left.\begin{aligned}\chi_r^2 &= \chi_a^2 + \sum_s \left[\frac{\left(\sum_t Y_{st}\right)^2}{n_s}\right] - \frac{\left(\sum_s \sum_t Y_{st}\right)^2}{\sum_s n_s} \\ &\qquad - \frac{\left\{\sum_s \left[(X_s - \bar{X})\left(\sum_t Y_{st}\right)\right]\right\}^2}{\sum_s [n_s(X_s - \bar{X})^2]} \\ &= \chi_a^2 + \chi_b^2, \text{ say}\end{aligned}\right] \qquad (11.61)$$

χ_b^2 is observed as equal to the "between means of arrays" sum of squares minus the quantity l, where

$$l = \frac{\left\{\sum_s \left[(X_s - \bar{X})\left(\sum_t Y_{st}\right)\right]\right\}^2}{\sum_s [n_s(X_s - \bar{X})^2]} \qquad (11.62)$$

We now test the hypothesis H_1 by calculating

$$F = \frac{\dfrac{\chi_b^2}{n_1}}{\dfrac{\chi_a^2}{n_2}} \qquad (11.63)$$

and then refer to Snedecor's tables of F (Table IV, Appendix) with $n = k - 2$ and $n_2 = \sum_s n_s - k$.

The components with the corresponding calculated values are then entered in an analysis-of-variance table. The quantity l is entered as the variation "due to linear regression" and χ_b^2 as the variation "due to departure from linear regression."

We now proceed to calculate l using the values recorded in Table 70, from which we get

$$l = \frac{(452.70)^2}{7046.68} = 29.08$$

Finally, the complete analysis in our problem is summarized in Table 71. For the test of the hypothesis H_1, we obtain

$$F_0 = \frac{87.046}{16.904} = 5.15$$

We enter the F-tables with $n_1 = 9$ and $n_2 = 252$ and find that F_0 is greater than the interpolated value of F at the 1 per cent level. Therefore, we reject the hypothesis H_1 and conclude that the regression of Y on X is nonlinear in form.

TABLE 70
CALCULATION OF THE VALUE OF l

n_s	X_s	$X_s - \bar{X}$	$\sum_t Y_{st}$	$(X_s - \bar{X})(\Sigma Y_{st})$	$n_s(X_s - \bar{X})^2$
12	10.5	*−10.6*	318.0	*−3370.80*	1348.32
10	12.5	*− 8.6*	265.0	*−2279.00*	739.60
14	14.5	*− 6.6*	407.0	*−2686.20*	609.84
30	16.5	*− 4.6*	841.0	*−3868.60*	634.80
41	18.5	*− 2.6*	1190.5	*−3095.30*	277.16
33	20.5	*− 0.6*	1016.5	*− 609.90*	11.88
29	22.5	1.4	852.5	1193.50	56.84
30	24.5	3.4	919.0	1286.60	346.80
32	26.5	5.4	1028.0	5551.20	933.12
22	28.5	7.4	713.0	5276.20	1204.72
10	30.5	9.4	325.0	3055.00	883.60
263	$\bar{X}$ = 21.1	......	......	452.70	7046.68

TABLE 71
ANALYSIS OF VARIANCE OF SCORES ON FINAL TEST—COMPLETE ANALYSIS

Source of variation	D.F.	Sum of squares	Mean square
Linear regression	1	29.08	29.080
Departure from linear regression	9	783.41	87.046
Within arrays	252	5950.21	16.904
Total	262	6762.70	......

$$F_0 = \frac{87.046}{16.904} = 5.15$$

The same methods could be used in testing the form of regression of X on Y.

Problem XI.5. The complete procedures for the analysis of variance and covariance for the data of a single classification. In order to illustrate how to calculate all the numerical values needed in a complete analysis of variance and covariance in the case of a single criterion of classification, how to proceed with the application of principles including the testing of underlying assumptions, and how to interpret the results, application has been made to the following problem. We wish to systematize the operations involved in the analysis in the most efficient way.

The primary data are given in Table 72, which gives the initial and final scores on a test of educational development, and the mental ages of 54 high-school students classified by grades, 18 students in each of the tenth, eleventh, and twelfth grades.

We wish to test the hypothesis that educational development is independent of grade, that is, that the mean achievements of students in the three grades are equal. The complete analysis is in three parts:

In Part I, we calculate all the values required for the complete analysis and carry out the analysis of variance on the final test scores;

In Part II, we give the complete procedures for the analysis of variance and covariance with one independent variable;

In Part III, we present the complete procedures for the analysis of variance and covariance with two independent variables.

TABLE 72

MENTAL AGES, INITIAL AND FINAL SCORES ON AN EDUCATIONAL DEVELOPMENT TEST OF 54 HIGH-SCHOOL STUDENTS CLASSIFIED BY GRADE*

Grade 10			Grade 11			Grade 12		
Final Y_{1t}	M.A. X_{1t}	Initial Z_{1t}	Final Y_{2t}	M.A. X_{2t}	Initial Z_{2t}	Final Y_{3t}	M.A. X_{3t}	Initial Z_{3t}
30	45	28	26	62	22	29	60	25
25	58	22	26	57	21	29	88	24
22	46	19	24	65	21	22	64	19
26	56	22	24	54	25	23	64	21
17	19	14	23	55	18	20	47	17
14	29	14	15	24	13	19	75	17
18	34	18	18	40	17	17	29	16
17	17	14	16	24	13	15	38	15
12	19	9	13	23	12	14	28	12
21	44	16	26	60	22	33	94	29
21	44	21	25	57	22	29	89	29
19	6	17	23	52	19	25	78	22
20	38	18	22	54	19	23	50	21
18	27	16	21	54	19	18	57	19
14	18	14	17	52	16	17	43	17
14	18	9	19	40	17	15	36	13
12	18	7	15	28	12	15	35	14
9	5	7	13	48	12	10	14	9

* Mental age, in terms of months, has been reduced by 100. Define Y_{st}, X_{st}, and Z_{st} as the final, mental age, and initial scores, respectively, for the tth individual in the sth group; where $s = 1, 2, 3$, denoting grade 10, 11, 12, respectively, and $t = 1, 2, \ldots, 18$.

PART I

Step 1. Calculate the following values:

(Some of the values reported here were calculated for later use and need not be considered in the analysis-of-variance procedure.)

$$a_{11} = \sum_t Y_{1t}^2 = 900 + \cdots + 81 = 6511$$

$$a_{21} = \sum_t Y_{2t}^2 = 676 + \cdots + 169 = 7806$$

$$a_{31} = \sum_t Y_{3t}^2 = 841 + \cdots + 100 = 8413$$

$$a_{12} = \sum_t X_{1t}^2 = 2025 + \cdots + 25 = 20{,}727$$

$$a_{22} = \sum_t X_{2t}^2 = 3844 + \cdots + 2304 = 43{,}317$$

$$a_{32} = \sum_t X_{3t}^2 = 3600 + \cdots + 196 = 63{,}595$$

$$a_{13} = \sum_t Z_{1t}^2 = 784 + \cdots + 49 = 5047$$

$$a_{23} = \sum_t Z_{2t}^2 = 484 + \cdots + 144 = 5970$$

$$a_{33} = \sum_t Z_{3t}^2 = 625 + \cdots + 81 = 6909$$

$$a_{14} = \sum_t (Y_{1t}X_{1t}) = 1350 + \cdots + 45 = 11{,}099$$

$$a_{24} = \sum_t (Y_{2t}X_{2t}) = 1612 + \cdots + 624 = 18{,}169$$

$$a_{34} = \sum_t (Y_{3t}X_{3t}) = 1740 + \cdots + 140 = 22{,}737$$

$$a_{15} = \sum_t (Y_{1t}Z_{1t}) = 840 + \cdots + 63 = 5701$$

$$a_{25} = \sum_t (Y_{2t}Z_{2t}) = 572 + \cdots + 156 = 6808$$

$$a_{35} = \sum_t (Y_{3t}Z_{3t}) = 725 + \cdots + 90 = 7607$$

$$a_{16} = \sum_t (X_{1t}Z_{1t}) = 1260 + \cdots + 35 = 9756$$

$$a_{26} = \sum_t (X_{2t}Z_{2t}) = 1364 + \cdots + 576 = 15{,}884$$

$$a_{36} = \sum_t (X_{3t}Z_{3t}) = 1500 + \cdots + 126 = 20{,}587$$

$$c_{11} = \frac{\left(\sum_t Y_{1t}\right)^2}{18} = \frac{(329)^2}{18} = 6013$$

$$c_{21} = \frac{\left(\sum_t Y_{2t}\right)^2}{18} = \frac{(366)^2}{18} = 7442$$

$$c_{31} = \frac{\left(\sum_t Y_{3t}\right)^2}{18} = \frac{(373)^2}{18} = 7729$$

$$c_{12} = \frac{\left(\sum_t X_{1t}\right)^2}{18} = \frac{(541)^2}{18} = 16{,}260$$

$$c_{22} = \frac{\left(\sum_t X_{2t}\right)^2}{18} = \frac{(849)^2}{18} = 40{,}045$$

$$c_{32} = \frac{\left(\sum_t X_{3t}\right)^2}{18} = \frac{(989)^2}{18} = 54{,}340$$

$$c_{13} = \frac{\left(\sum_t Z_{1t}\right)^2}{18} = \frac{(285)^2}{18} = 4513$$

$$c_{23} = \frac{\left(\sum_t Z_{2t}\right)^2}{18} = \frac{(320)^2}{18} = 5689$$

$$c_{33} = \frac{\left(\sum_t Z_{3t}\right)^2}{18} = \frac{(339)^2}{18} = 6385$$

$$c_{14} = \frac{\left(\sum_t Y_{1t}\right)\left(\sum_t X_{1t}\right)}{18} = \frac{329(541)}{18} = 9888$$

$$c_{24} = \frac{\left(\sum_t Y_{2t}\right)\left(\sum_t X_{2t}\right)}{18} = \frac{366(849)}{18} = 17{,}263$$

$$c_{34} = \frac{\left(\sum_t Y_{3t}\right)\left(\sum_t X_{3t}\right)}{18} = \frac{373(989)}{18} = 20{,}494$$

$$c_{15} = \frac{\left(\sum_t Y_{1t}\right)\left(\sum_t Z_{1t}\right)}{18} = \frac{329(285)}{18} = 5209$$

$$c_{25} = \frac{\left(\sum_t Y_{2t}\right)\left(\sum_t Z_{2t}\right)}{18} = \frac{366(320)}{18} = 6507$$

$$c_{35} = \frac{\left(\sum_t Y_{3t}\right)\left(\sum_t Z_{3t}\right)}{18} = \frac{373(339)}{18} = 7025$$

$$c_{16} = \frac{\left(\sum_t X_{1t}\right)\left(\sum_t Z_{1t}\right)}{18} = \frac{541(285)}{18} = 8566$$

$$c_{26} = \frac{\left(\sum_t X_{2t}\right)\left(\sum_t Z_{2t}\right)}{18} = \frac{849(320)}{18} = 15{,}093$$

$$c_{36} = \frac{\left(\sum_t X_{3t}\right)\left(\sum_t Z_{3t}\right)}{18} = \frac{989(339)}{18} = 18{,}626$$

$$d_1 = \frac{\left(\sum_s \sum_t Y_{st}\right)^2}{54} = \frac{(1068)^2}{54} = 21{,}123$$

$$d_2 = \frac{\left(\sum_s \sum_t X_{st}\right)^2}{54} = \frac{(2379)^2}{54} = 104{,}808$$

$$d_3 = \frac{\left(\sum_s \sum_t Z_{st}\right)^2}{54} = \frac{(944)^2}{54} = 16{,}503$$

$$d_4 = \frac{\left(\sum_s \sum_t Y_{st}\right)\left(\sum_s \sum_t X_{st}\right)}{54} = \frac{1068(2379)}{54} = 47{,}051$$

$$d_5 = \frac{\left(\sum_s \sum_t Y_{st}\right)\left(\sum_s \sum_t Z_{st}\right)}{54} = \frac{1068(944)}{54} = 18{,}670$$

$$d_6 = \frac{\left(\sum_s \sum_t X_{st}\right)\left(\sum_s \sum_t Z_{st}\right)}{54} = \frac{2379(944)}{54} = 41{,}588$$

$$a_1 = \sum_s \sum_t Y_{st}^2 = a_{11} + a_{21} + a_{31} = 22{,}730$$

$$a_2 = \sum_s \sum_t X_{st}^2 = a_{12} + a_{22} + a_{32} = 127{,}639$$

$$a_3 = \sum_s \sum_t Z_{st}^2 = a_{13} + a_{23} + a_{33} = 17{,}926$$

$$a_4 = \sum_s \sum_t (Y_{st}X_{st}) = a_{14} + a_{24} + a_{34} = 52{,}005$$

$$a_5 = \sum_s \sum_t (Y_{st}Z_{st}) = a_{15} + a_{25} + a_{35} = 20{,}116$$

$$a_6 = \sum_s \sum_t (X_{st}Z_{st}) = a_{16} + a_{26} + a_{36} = 46{,}227$$

$$c_1 = \frac{\sum_s \left(\sum_t Y_{st}\right)^2}{18} = c_{11} + c_{21} + c_{31} = 21{,}184$$

$$c_2 = \frac{\sum_s \left(\sum_t X_{st}\right)^2}{18} = c_{12} + c_{22} + c_{32} = 110{,}645$$

$$c_3 = \frac{\sum\limits_s \left(\sum\limits_t Z_{st}\right)^2}{18} = c_{13} + c_{23} + c_{33} = 16{,}587$$

$$c_4 = \frac{\sum\limits_s \left[\left(\sum\limits_t Y_{st}\right)\left(\sum\limits_t X_{st}\right)\right]}{18} = c_{14} + c_{24} + c_{34} = 47{,}645$$

$$c_5 = \frac{\sum\limits_s \left[\left(\sum\limits_t Y_{st}\right)\left(\sum\limits_t Z_{st}\right)\right]}{18} = c_{15} + c_{25} + c_{35} = 18{,}741$$

$$c_6 = \frac{\sum\limits_s \left[\left(\sum\limits_t X_{st}\right)\left(\sum\limits_t Z_{st}\right)\right]}{18} = c_{16} + c_{26} + c_{36} = 42{,}285$$

Step 2. Calculate the sum of squares of y for each group. Define

$$\theta_s = \sum_t (Y_{st} - \bar{Y}_s)^2 = \sum_t y_{st}^2$$

where $\bar{Y}_s = \frac{\sum\limits_t Y_{st}}{18}$

It is obvious that

$$\theta_s = \sum_t Y_{st}^2 - \frac{\left(\sum\limits_t Y_{st}\right)^2}{18}$$

Therefore, we have

$$\theta_1 = a_{11} - c_{11} = 498 = \sum_t y_{1t}^2$$

$$\theta_2 = a_{21} - c_{21} = 364 = \sum_t y_{2t}^2$$

$$\theta_3 = a_{31} - c_{31} = 684 = \sum_t y_{3t}^2$$

Step 3. Use the L_1-criterion to test the hypothesis $H_1 : \sigma_s = \sigma$. The calculations involved are summarized in Table 73.

Step 4. Calculate the following values for the analysis of variance of y. The sums of squares for the different sources of variation are (see Step 1):

(1) Within grades = $a_1 - c_1 = 1546$
(2) Between grades = $c_1 - d_1 = 61$
(3) Total = $a_1 - d_1 = 1607$

TABLE 73

L_1-CALCULATIONS FOR $H_1: \sigma_s = \sigma$

f_s	n_s	$\log n_s$	$n_s \log n_s$	θ_s'	$\log \theta_s'$	$n_s \log \theta_s'$
17	18	1.2553	...	498	2.6972	...
17	18	1.2553	...	364	2.5611	...
17	18	1.2553	...	684	2.8351	...
51	54	$\sum_s n_s \log n_s = 67.7862$		1546	$\sum_s n_s \log \theta_s = 145.6812$	

$$\begin{aligned}\log L_1 &= \log N - \frac{1}{N}\sum_s n_s \log n_s + \frac{1}{N}\sum_s n_s \log \theta_s' - \log\left(\sum_s \theta_s'\right)\\ &= \log 54 - \tfrac{1}{54}\,(67.7862) + \tfrac{1}{54}(145.6812) - \log 1546\\ &= 1.7324 - 1.2553 + 2.6978 - 3.1892\\ &= 9.9857 - 10\\ \therefore L_1 &= .968\end{aligned}$$

Refer to Nayer's tables of L_1 (Table V, Appendix) with $k = 3$ and degrees of freedom $f = 17$. We have $P > .05$. Therefore we accept H_1. Assuming that the three groups have common variance, we may combine the results.

Step 5. Analysis of variance to test the hypothesis $H_0: \bar{Y}_s = \bar{Y}$. The results are summarized in Table 74.

TABLE 74

ANALYSIS OF VARIANCE OF FINAL SCORE FOR DIFFERENT GRADES

Source of variation	D.F.	Sum of squares	Mean square	F	Hypothesis tested
Within grades	51	1546	30.31	....	....
Between grades	2	61	30.50	1.01	Accepted
Total	53	1607			

Where $F = \dfrac{\text{mean square of between grades}}{\text{mean square of within grades}}$

Refer to Snedecor's tables of F (Table IV, Appendix) with $n_1 = 2$ and $n_2 = 51$. We have $P > .05$. Therefore, we accept the hypothesis H_0 and conclude that there are no significant differences among the means of the three grades.

PART II

Complete procedures for the analysis of variance and covariance with one independent variable.

Step 1. Calculate the following values (see Part I, Steps 1 and 2):

$$\sum_t x_{1t}^2 = \sum_t X_{1t}^2 - \frac{\left(\sum_t X_{1t}\right)^2}{18} = a_{12} - c_{12} = 4467$$

$$\sum_t x_{2t}^2 = \sum_t X_{2t}^2 - \frac{\left(\sum_t X_{2t}\right)^2}{18} = a_{22} - c_{22} = 3272$$

$$\sum_t x_{3t}^2 = \sum_t X_{3t}^2 - \frac{\left(\sum_t X_{3t}\right)^2}{18} = a_{32} - c_{32} = 9255$$

$$\sum_t (y_{1t}x_{1t}) = \sum_t (Y_{1t}X_{1t}) - \frac{\left(\sum_t Y_{1t}\right)\left(\sum_t X_{1t}\right)}{18} = a_{14} - c_{14} = 1211$$

$$\sum_t (y_{2t}x_{2t}) = \sum_t (Y_{2t}X_{2t}) - \frac{\left(\sum_t Y_{2t}\right)\left(\sum_t X_{2t}\right)}{18} = a_{24} - c_{24} = 906$$

$$\sum_t (y_{3t}x_{3t}) = \sum_t (Y_{3t}X_{3t}) - \frac{\left(\sum_t Y_{3t}\right)\left(\sum_t X_{3t}\right)}{18} = a_{34} - c_{34} = 2243$$

From Part I, Step 2, we get

$$\sum_t y_{1t}^2 = 498$$

$$\sum_t y_{2t}^2 = 364$$

$$\sum_t y_{3t}^2 = 684$$

Then, we have

$$M_1 = \frac{\left[\sum_t (y_{1t}x_{1t})\right]^2}{\sum_t x_{1t}^2} = \frac{(1211)^2}{4467} = 328$$

$$M_2 = \frac{\left[\sum_t (y_{2t}x_{2t})\right]^2}{\sum_t x_{2t}^2} = \frac{(906)^2}{3272} = 251$$

$$M_3 = \frac{\left[\sum_t (y_{3t}x_{2t})\right]^2}{\sum_t x_{3t}^2} = \frac{(2243)^2}{9255} = 544$$

Define

$$\text{Adjusted} \sum_t y_{st}^2 = \sum_t (y_{st} - b_s x_{st})^2$$

where $b_s = \dfrac{\sum\limits_t (y_{st}x_{st})}{\sum\limits_t x^2_{st}}$

By simple algebraic operation, we have

$$\text{Adjusted} \sum_t y^2_{st} = \sum_t y^2_{st} - \frac{\left[\sum\limits_t (y_{st}x_{st})\right]^2}{\sum\limits_t x^2_{st}} = \sum_t y^2_{st} - M_s$$

Define

$$\theta^1_s = \text{adjusted} \sum_t y^2_{st}$$

Therefore, we have

$$\theta^1_1 = \sum_t y^2_{1t} - M_1 = 170$$

$$\theta^1_2 = \sum_t y^2_{2t} - M_2 = 113$$

$$\theta^1_3 = \sum_t y^2_{3t} - M_3 = 140$$

Step 2. Use the L_1-criterion to test the hypothesis $H_1' : \sigma_{y_s \cdot x_s} = \sigma_{y \cdot x}$ The calculations involved are summarized in Table 75.

TABLE 75
L_1-CALCULATIONS FOR $H_1' : \sigma_{y_s \cdot x_s} = \sigma_{y \cdot x}$

f_s	n_s	$\log n_s$	$n_s \log n_s$	θ_s'	$\log \theta_s'$	$n_s \log \theta_s'$
16	18	1.2553	...	170	2.2305	...
16	18	1.2553	...	113	2.0531	...
16	18	1.2553	...	140	2.1461	...
48	54	$\sum\limits_s n_s \log n_s = 67.7862$		423	$\sum\limits_s n_s \log \theta_s' = 115.7346$	

$$\log L_1 = \log N - \frac{1}{N}\sum_s n_s \log n_s + \frac{1}{N}\sum_s n_s \log \theta_s' - \log \left(\sum_s \theta_s'\right)$$
$$= \log 54 - \tfrac{1}{54}(67.7862) + \tfrac{1}{54}(115.7346) - \log 423$$
$$= 1.7324 - 1.2553 + 2.1432 - 2.6263 = 9.9940 - 10$$
$$\therefore L_1 = .986$$

Refer to Nayer's tables of L_1 with $k = 3$ and degrees of freedom $f = 16$. We have $P > .05$. Therefore, we accept H_1' and combine the results.

Step 3. Calculate the following values for the analysis of variance of y and x and the covariance of yx (with x held constant). The sums of squares and of products for the different sources of variation are (see Part I, Step 1):

$$\text{(1) Within grades:} \left\{ \begin{aligned} \Sigma y^2 &= a_1 - c_1 = 1546 = A_0 \\ \Sigma x^2 &= a_2 - c_2 = 16{,}994 = B_0 \\ \Sigma yx &= a_4 - c_4 = 4360 = D_0 \end{aligned} \right.$$

$$\text{(2) Between grades:} \left\{ \begin{aligned} \Sigma y^2 &= c_1 - d_1 = 61 = A_1 \\ \Sigma x^2 &= c_2 - d_2 = 5837 = B_1 \\ \Sigma yx &= c_4 - d_4 = 594 = D_1 \end{aligned} \right.$$

$$\text{(3) Total:} \left\{ \begin{aligned} \Sigma y^2 &= a_1 - d_1 = 1607 = A \\ \Sigma x^2 &= a_2 - d_2 = 22{,}831 = B \\ \Sigma yx &= a_4 - d_4 = 4954 = D \end{aligned} \right.$$

Step 4. Calculate $b\Sigma yx$ for "within" and "total" where $b\Sigma yx = \dfrac{(\Sigma yx)^2}{\Sigma x^2}$.

Refer to Step 3; we have

$$\text{(1) Within grades: } b\Sigma yx = \frac{D_0^2}{B_0} = 1119 = M_0$$

$$\text{(2) Total: } b\Sigma yx = \frac{D^2}{B} = 1075 = M$$

Step 5. Calculate adjusted Σy^2 for "within" and "total," and reduced Σy^2 for "between."

(1) Within grades: Adjusted $\Sigma y^2 = A_0 - M_0 = 427 = P_0$
(2) Total: Adjusted $\Sigma y^2 = A - M = 532 = P$
(3) Between grades: Reduced $\Sigma y^2 = P - P_0 = 105$

Step 6. Analysis of variance and covariance to test the hypothesis $H_0^1: \bar{Y}_s = \bar{Y}$ with X held constant. The results are summarized in Table 76.

TABLE 76
ANALYSIS OF VARIANCE AND COVARIANCE OF FINAL SCORE WITH MENTAL AGE HELD CONSTANT

Source of variation	D.F.	Σy^2	Σx^2	Σxy	Adjusted or reduced				
					D.F.	S.S.	M.S.	F	Hypothesis
Within grades	51	1546	16,994	4360	50	427	8.54	....	....
Between grades	2	61	5,837	594	2	105	52.50	6.148	Rejected
Total	53	1607	22,831	4954	52	532			

Refer to Snedecor's tables of F with $n_1 = 2$ and $n_2 = 50$. We have $P < .01$. Therefore, we reject the hypothesis $H_0{}^1$ and conclude that there are significant differences among the means of final scores for these three grades with the effects of mental age partialed out.

PART III

Complete procedures for the analysis of variance and covariance with two independent variables.

Step 1. Calculate the following values (see Part I, Steps 1 and 2):

$$\sum_t z_{1t}^2 = \sum_t Z_{1t}^2 - \frac{\left(\sum_t Z_{1t}\right)^2}{18} = a_{13} - c_{13} = 534$$

$$\sum_t z_{2t}^2 = \sum_t Z_{2t}^2 - \frac{\left(\sum_t Z_{2t}\right)^2}{18} = a_{23} - c_{23} = 281$$

$$\sum_t z_{3t}^2 = \sum_t Z_{3t}^2 - \frac{\left(\sum_t Z_{3t}\right)^2}{18} = a_{33} - c_{33} = 524$$

$$\sum_t (y_{1t}z_{1t}) = \sum_t (Y_{1t}Z_{1t}) - \frac{\left(\sum_t Y_{1t}\right)\left(\sum_t Z_{1t}\right)}{18} = a_{15} - c_{15} = 492$$

$$\sum_t (y_{2t}z_{2t}) = \sum_t (Y_{2t}Z_{2t}) - \frac{\left(\sum_t Y_{2t}\right)\left(\sum_t Z_{2t}\right)}{18} = a_{25} - c_{25} = 301$$

$$\sum_t (y_{3t}z_{3t}) = \sum_t (Y_{3t}Z_{3t}) - \frac{\left(\sum_t Y_{3t}\right)\left(\sum_t Z_{3t}\right)}{18} = a_{35} - c_{35} = 582$$

$$\sum_t (x_{1t}z_{1t}) = \sum_t (X_{1t}Z_{1t}) - \frac{\left(\sum_t X_{1t}\right)\left(\sum_t Z_{1t}\right)}{18} = a_{16} - c_{16} = 1190$$

$$\sum_t (x_{2t}z_{2t}) = \sum_t (X_{2t}Z_{2t}) - \frac{\left(\sum_t X_{2t}\right)\left(\sum_t Z_{2t}\right)}{18} = a_{26} - c_{26} = 751$$

$$\sum_t (x_{3t}z_{3t}) = \sum_t (X_{3t}Z_{3t}) - \frac{\left(\sum_t X_{3t}\right)\left(\sum_t Z_{3t}\right)}{18} = a_{36} - c_{36} = 1961$$

From Part I, Step 2, and Part II, Step 1, we also have

$$\sum_t y_{1t}^2 = 498$$

$$\sum_t y_{2t}^2 = 364$$

$$\sum_t y_{3t}^2 = 684$$

$$\sum x_{1t}^2 = 4467$$

$$\sum_t x_{2t}^2 = 3272$$

$$\sum_t x_{3t}^2 = 9255$$

$$\sum_t (y_{1t}x_{1t}) = 1211$$

$$\sum_t (y_{2t}x_{2t}) = 906$$

$$\sum_t (y_{3t}x_{3t}) = 2243$$

Then, we have

$$M_1^1 = \frac{\sum_t z_{1t}^2 \left[\sum_t (y_{1t}x_{1t})\right]^2 - \sum_t (x_{1t}z_{1t}) \sum_t (y_{1t}z_{1t}) \sum_t (y_{1t}x_{1t})}{\sum_t x_{1t}^2 \sum_t z_{1t}^2 - \left[\sum_t (x_{1t}z_{1t})\right]^2}$$

$$= \frac{534(1211)^2 - 1190(492)(1211)}{4467(534) - (1190)^2}$$

$$= \frac{783{,}122{,}214 - 709{,}016{,}280}{969{,}278} = \frac{74{,}105{,}934}{969{,}278} = 76$$

$$M_2^1 = \frac{\sum_t z_{2t}^2 \left[\sum_t (y_{2t}x_{2t})\right]^2 - \sum_t (x_{2t}z_{2t}) \sum_t (y_{2t}z_{2t}) \sum_t (y_{2t}x_{2t})}{\sum_t x_{2t}^2 \sum_t z_{2t}^2 - \left[\sum_t (x_{2t}z_{2t})\right]^2}$$

$$= \frac{281(906)^2 - 751(301)(906)}{3272(281) - (751)^2}$$

$$= \frac{230{,}654{,}916 - 204{,}802{,}206}{355{,}431} = \frac{25{,}852{,}710}{355{,}431} = 73$$

$$M_3^1 = \frac{\sum_t z_{3t}^2 \left[\sum_t (y_{3t}x_{3t})\right]^2 - \sum_t (x_{3t}z_{3t}) \sum_t (y_{3t}z_{3t}) \sum_t (y_{3t}x_{3t})}{\sum_t x_{3t}^2 \sum_t z_{3t}^2 - \left[\sum_t (x_{3t}z_{3t})\right]^2}$$

$$= \frac{524(2243)^2 - 1961(582)(2243)}{9255(524) - (1961)^2}$$

$$= \frac{2{,}636{,}269{,}676 - 2{,}559{,}940{,}386}{1{,}004{,}099} = \frac{76{,}329{,}290}{1{,}004{,}099} = 76$$

$$N_1^1 = \frac{\sum_t x_{1t}^2 \left[\sum_t (y_{1t}z_{1t})\right]^2 - \sum_t (x_{1t}z_{1t}) \sum_t (y_{1t}x_{1t}) \sum_t (y_{1t}z_{1t})}{\sum_t x_{1t}^2 \sum_t z_{1t}^2 - \left[\sum_t (x_{1t}z_{1t})\right]^2}$$

$$= \frac{4467(492)^2 - 709{,}016{,}280}{969{,}278}$$

$$= \frac{1{,}081{,}299{,}888 - 709{,}016{,}280}{969{,}278} = \frac{372{,}283{,}608}{969{,}278} = 384$$

$$N_2^1 = \frac{\sum_t x_{2t}^2 \left[\sum_t (y_{2t}z_{2t})\right]^2 - \sum_t (x_{2t}z_{2t}) \sum_t (y_{2t}x_{2t}) \sum_t (y_{2t}z_{2t})}{\sum_t x_{2t}^2 \sum_t z_{2t}^2 - \left[\sum_t (x_{2t}z_{2t})\right]^2}$$

$$= \frac{3272(301)^2 - 204{,}802{,}206}{355{,}431}$$

$$= \frac{297{,}431{,}344 - 204{,}802{,}206}{355{,}431} = \frac{92{,}629{,}138}{355{,}431} = 261$$

$$N_3^1 = \frac{\sum_t x_{3t}^2 \left[\sum_t (y_{3t}z_{3t})\right]^2 - \sum_t (x_{3t}z_{3t}) \sum_t (y_{3t}x_{3t}) \sum_t (y_{3t}z_{3t})}{\sum_t x_{3t}^2 \sum_t z_{3t}^2 - \left[\sum_t (x_{3t}z_{3t})\right]^2}$$

$$= \frac{9255(582)^2 - 2{,}559{,}940{,}386}{1{,}004{,}099}$$

$$= \frac{3{,}134{,}890{,}620 - 2{,}559{,}940{,}386}{1{,}004{,}099} = \frac{574{,}950{,}234}{1{,}004{,}099} = 573$$

Define

$$\text{Adjusted} \sum_t y_{st}^2 = \sum_t (y_{st} - b_{s1}x_{st} - b_{s2}z_{st})^2$$

$$\text{where } b_{s1} = \frac{\sum_t z_{st}^2 \sum_t (y_{st}x_{st}) - \sum_t (x_{st}z_{st}) \sum_t (y_{st}z_{st})}{\sum_t x_{st}^2 \sum_t z_{st}^2 - \left[\sum_t (x_{st}z_{st})\right]^2}$$

$$b_{s2} = \frac{\sum_t x_{st}^2 \sum_t (y_{st}z_{st}) - \sum_t (x_{st}z_{st}) \sum_t (y_{st}x_{st})}{\sum_t x_{st}^2 \sum_t z_{st}^2 - \left[\sum_t (x_{st}z_{st})\right]^2}$$

By troublesome algebraic operation, we have

$$\text{Adjusted} \sum_t y_{st}^2 = \sum_t y_{st}^2 - M_s^1 - N_s^1$$

Define $\theta_s'' = \text{adjusted} \sum_t y_{st}^2$

Therefore, we have

$$\theta_1'' = \sum_t y_{1t}^2 - M_1^1 - N_1^1 = 38$$

$$\theta_2'' = \sum_t y_{2t}^2 - M_2^1 - N_2^1 = 30$$

$$\theta_3'' = \sum y_{3t}^2 - M_3^1 - N_3^1 = 35$$

Step 2. Use the L_1-criterion to test the hypothesis $H_1'': \sigma_{y_s \cdot x_s z_s} = \sigma_{y \cdot xz}$. The calculations involved are summarized in Table 77.

TABLE 77
L_1-CALCULATIONS FOR $H_1'': \sigma_{y_s \cdot x_s z_s} = \sigma_{y \cdot xz}$

f_s	n_s	$\log n_s$	$n_s \log n_s$	θ_s''	$\log \theta_s''$	$n_s \log \theta_s'$
15	18	1.2553	...	38	1.5798	...
15	18	1.2553	...	30	1.4771	...
15	18	1.2553	...	35	1.5441	...
45	54	$\sum_s n_s \log n_s = 67.7862$		103	$\sum_s n_s \log \theta_s'' = 82.8180$	

$$\begin{aligned} \log L_1 &= \log N - \frac{1}{N}\sum_s n_s \log n_s + \frac{1}{N}\sum_s n_s \log \theta_s'' - \log\left(\sum_s \theta_s''\right) \\ &= \log 54 - \tfrac{1}{54}(67.7862) + \tfrac{1}{54}(82.8180) - \log 106 \\ &= 1.7324 - 1.2553 + 1.5337 - 2.0128 = 9.9980 - 10 \\ L_1 &= .995 \end{aligned}$$

Refer to Nayer's tables of L_1 with $k = 3$ and degrees of freedom $f = 15$. We have $P > .05$. Therefore, we accept H_1'' and combine the results.

Step 3. Calculate the necessary values for the analysis of variance of x, y, and z and covariance of yx, yz, and xz (with both x and z held constant). The sums of squares and of products for the different sources of variation are (see Part I, Step 1 and Part II, Step 3):

$$\text{(1) Within grades} \left\{ \begin{aligned} \Sigma y^2 &= 1546 = A_0 \\ \Sigma x^2 &= 16{,}994 = B_0 \\ \Sigma z^2 &= a_3 - c_3 = 1339 = C_0 \\ \Sigma yx &= 4360 = D_0 \\ \Sigma yz &= a_5 - c_5 = 1375 = E_0 \\ \Sigma xz &= a_6 - c_6 = 3942 = F_0 \end{aligned} \right.$$

$$\text{(2) Between grades} \left\{ \begin{aligned} \Sigma y^2 &= 61 = A_1 \\ \Sigma x^2 &= 5837 = B_1 \\ \Sigma z^2 &= c_3 - d_3 = 84 = C_1 \\ \Sigma yx &= 594 = D_1 \\ \Sigma yz &= c_5 - d_5 = 71 = E_1 \\ \Sigma xz &= c_6 - d_6 = 697 = F_1 \end{aligned} \right.$$

$$\text{(3) Total} \left\{ \begin{aligned} \Sigma y^2 &= 1607 = A \\ \Sigma x^2 &= 22{,}831 = B \\ \Sigma z^2 &= a_3 - d_3 = 1423 = C \\ \Sigma yx &= 4954 = D \\ \Sigma yz &= a_5 - d_5 = 1446 = E \\ \Sigma xz &= a_6 - d_6 = 4639 = F \end{aligned} \right.$$

Step 4. Calculate $b_1\Sigma yx$ and $b_2\Sigma yz$ for "within" and "total,"

where
$$b_1\sum yx = \frac{\Sigma z^2(\Sigma yx)^2\Sigma xz\Sigma yz\Sigma yx}{\Sigma x^2\Sigma z^2 - (\Sigma xz)^2}$$
$$b_2\sum yz = \frac{\Sigma x^2(\Sigma yz)^2\Sigma xz\Sigma yx\Sigma yz}{\Sigma x^2\Sigma z^2 - (\Sigma xz)^2}$$

Refer to Step 3. We have

(1) Within grades: $$b_1\sum yx = \frac{C_0D_0^2 - F_0E_0D_0}{B_0C_0 - F_0^2} = 252 = M_0^1$$
$$b_2\sum yz = \frac{B_0E_0^2 - F_0D_0E_0}{B_0C_0 - F_0^2} = 1178 = N_0^1$$

Total: $$b_1\sum yx = \frac{CD^2 - FED}{BC - F^2} = 154 = M^1$$
$$b_2\sum yz = \frac{BE^2 - FDE}{BC - F^2} = 1323 = N^1$$

Step 5. Calculate adjusted Σy^2 for "within" and "total," and reduced Σy^2 for "between."

(1) Within grades: adjusted $\Sigma y^2 = A_0 - M_0^1 - N_0^1 = 116 = P_0^1$
(2) Total: adjusted $\Sigma y^2 = A - M - N = 130 = P^1$
(3) Between grades: reduced $\Sigma y^2 = P^1 - P_0^1 = 14$

Step 6. Analysis of variance and covariance to test the hypothesis $H_0'': \bar{Y}_s = \bar{Y}$ with both X and Z held constant. The results are summarized in Table 78.

TABLE 78

ANALYSIS OF VARIANCE AND COVARIANCE OF FINAL SCORE WITH BOTH MENTAL AGE AND INITIAL SCORE HELD CONSTANT

Source of variation	D.-F.	Σy^2	Σx^2	Σz^2	Σyx	Σyz	Σxz	Adjusted or reduced				
								D.-F.	S.-S.	M.-S.	F	Hypothesis
Within grades	51	1,546	16,994	1,339	4,360	1,375	3,942	49	116	2.37	....	....
Between grades	2	61	5,837	84	594	71	697	2	14	7.00	2.95	Accepted
Total	53	1,607	22,831	1,423	4,954	1,446	4,639	51	130			

Refer to Snedecor's tables of F with $n_1 = 2$ and $n_2 = 49$. We have $P > .05$. So we accept the hypothesis H_0'' and conclude that there are no significant differences among the means of final scores for these three grades with both the effects of mental age and initial score partialed out.

Analysis of Variance in the Case of Unequal or Disproportionate Numbers of Observations in the Subclasses. The analysis of variance in the case of a single criterion of classification with unequal numbers in the subclasses introduces no new difficulty as has been indicated in Problem XI.2. However, when our data have been classified on the basis of two or more criteria with unequal subclass numbers, new difficulties arise.

In agriculture and in other experimental sciences it is usually possible to design an experiment so that each subclass has always the same number of individuals. If this were a necessary condition, the use of the powerful tool of analysis of variance would be greatly restricted, since there are fields, such as those dealing with human beings—education and psychology, for instance—where unequal representation in each cell of multiple classification of data is of common occurrence, both in experimentation and in other observational programs, including data collected by governmental and state agencies. There is an urgent need, therefore, for a systematic formulation of methods of attacking problems when unequal representation in the subclasses occurs. Methods have been developed for such problems (Refs. 8, 9, 10).

Tsao (Ref. 10) has treated the problem of analysis of variance and covariance for unequal or disproportionate representation in the subclasses by giving the mathematical solution with the specified restrictions defined and by proposing new approximate methods with the respective statistical assumptions to be fulfilled. Our consideration of this problem is limited to the presentation of an approximation method of analysis for unequal representation in the subclasses of two classifications.

Problem XI.6. An approximation method of analysis of variance for unequal frequencies in the subclasses of two classifications. We take the problem of testing two hypotheses: (1) that the grade means on a speed of reading test are equal and (2) that the school means on the reading test are equal. The basic data for the fifth, sixth, seventh, and eighth grades in each of two schools are given in Table 79, including the appropriate notations. The complete analysis of the problem follows.

TABLE 79
CALCULATED MEASURES FOR SPEED SCORE IN GATES READING-SURVEY TEST

School	Grade	n_{si}	$\bar{X}_{si}$	s'_{si}	$\sum_t x^2_{sit} = \sum_t (X_{sit} - \bar{X}_{si})^2$
A	5	41	49.68	12.53	6280
	6	39	41.08	11.28	4835
	7	32	42.41	9.99	3094
	8	36	53.25	10.59	3925
B	5	26	33.92	12.47	3888
	6	27	29.22	11.02	3157
	7	34	32.50	10.00	3300
	8	32	40.53	9.84	3002

where $s = 1, 2, 3, 4$ represent grades 5, 6, 7, and 8, respectively
$i = 1, 2$ represent schools A and B, respectively
$t = 1, 2, \cdots, n_{si}$
n_{si} is the number of observations for the sth grade in the ith school
$\bar{X}_{si}$ is the mean score for the sth grade in the ith school
s'_{si} is the unbiased estimate of standard deviation for the sth grade in the ith school
$\bar{X}_{si}$ and s'_{si} are obtained by the following definitions:

$$\bar{X}_{si} = \frac{\sum\limits_{t} X_{sit}}{n_{si}};$$

$$s'_{si} = \sqrt{\frac{\sum\limits_{t} x^2{}_{sit}}{n_{si} - 1}} = \sqrt{\frac{\sum\limits_{t} (X^2{}_{sit} - X_{si})^2}{n_{si} - 1}}$$

Step 1. Use criterion L_1 to test the hypothesis $H_1{:}\sigma_{si} = \sigma$.
Let us define:

$$\theta'_{si} = \sum_{t} x^2_{sit} = \sum_{t} (X_{sit} - \bar{X}_{si})^2$$

The calculations for L_1 are summarized in Table 80.

TABLE 80
L_1-CALCULATIONS FOR $H_1{:}\sigma_{si} = \sigma$

f_{si}	n_{si}	$\log n_{si}$	$n_{si} \log n_{si}$	θ'_{si}	$\log \theta'_{si}$	$n_{si} \log \theta'_{si}$
40	41	1.6128	....	6,280	3.7980	....
38	39	1.5911	....	4,835	3.6844	....
31	32	1.5051	....	3,094	3.4905	....
35	36	1.5563	....	3,925	3.5938	....
25	26	1.4150	....	3,888	3.5897	....
26	27	1.4314	....	3,157	3.4993	....
33	34	1.5315	....	3,300	3.5185	....
31	32	1.5051	....	3,002	3.4774	....
$N = 267$		$\sum\limits_{si} n_{si} \log n_{si} = 408.0397$		31,481	$\sum\limits_{si} n_{si} \log \theta'_{si} = 959.2015$	

The harmonic mean of f_{si}

$$= \frac{8}{\frac{1}{40} + \frac{1}{38} + \frac{1}{31} + \frac{1}{35} + \frac{1}{25} + \frac{1}{26} + \frac{1}{33} + \frac{1}{31}}$$

$$= \frac{8}{.253168} = 31.60$$

$$\begin{aligned}\log L_1 &= \log N - \frac{1}{N} \sum_{si} n_{si} \log n_{si} + \frac{1}{N} \sum_{si} n_{si} \log \theta'_{si} - \log \sum_{si} \theta'_{si} \\ &= \log 267 - \tfrac{1}{267}(408.0397) + \tfrac{1}{267}(959.2015) - \log 31{,}481 \\ &= 2.4265 - 1.5282 + 3.5936 - 4.4980 = 9.9939 - 10\end{aligned}$$

Therefore, $L_1 = .986$. Refer to Nayer's tables of L_1 (Table V, Appendix) with $k = 8$ and degrees of freedom $\hat{f} = 31.60$. We find that $P > .05$. Therefore, we accept H_1. We may assume that the eight groups have a common variance, and combine the results.

Step 2. Use the χ^2-criterion to test the goodness of fit for the equal frequencies in each subclass. First calculate the mean frequency: $\bar{n} = \frac{N}{8} = \frac{267}{8} = 33.375$. The results of the χ^2 test are summarized in Table 81.

TABLE 81
CALCULATION OF χ^2

f_0	f_t	$\lvert f_0 - f_t \rvert$	$(f_0 - f_t)^2$	$\frac{(f_0 - f_t)^2}{f_t}$
41	33.375	7.625	58.140625	1.7420
39	33.375	5.625	31.640625	0.9480
32	33.375	1.375	1.890625	0.0566
36	33.375	2.625	6.890625	0.2065
26	33.375	7.375	54.390625	1.6297
27	33.375	6.375	40.640625	1.2177
34	33.375	0.625	0.390625	0.0117
32	33.375	1.375	1.890625	0.0566
267	267.000			$\chi_0^2 = 5.8688$

We find

$$\chi_0^2 = 5.8688$$

Refer to χ^2-table with d.f. = 7. We find $.70 > P > .50$. Therefore, we conclude that for our data the class numbers do not differ significantly. It is justifiable to use the approximation method.

Step 3. Convert Table 79 into a table with equal frequency of 33.375 for each subclass.

Retaining the original estimate of the standard deviation, we have the following estimates of $\sum_t x_{sit}^2$:

$$\sum_t x_{11t}^2 = \frac{33.375}{41}(6280) = 5112 \qquad \sum_t x_{21t}^2 = \frac{33.375}{26}(3888) = 4991$$

$$\sum_t x_{12t}^2 = \frac{33.375}{39}(4835) = 4138 \qquad \sum_t x_{22t}^2 = \frac{33.375}{27}(3157) = 3902$$

$$\sum_t x_{13t}^2 = \frac{33.375}{32}(3094) = 3227 \qquad \sum_t x_{23t}^2 = \frac{33.375}{34}(3300) = 3239$$

$$\sum_t x_{14t}^2 = \frac{33.375}{36}(3925) = 3639 \qquad \sum_t x_{24t}^2 = \frac{33.375}{32}(3002) = 3131$$

These results, together with the data in Table 79, are summarized in Table 82. The notations are the same as in Table 79.

TABLE 82
EXPECTED MEASURES FOR SPEED SCORE IN GATES READING SURVEY

School	Grade (1)	n (2)	$\bar{X}_{si}$ (3)	S'_{si} (4)	$\sum_t x^2_{sit}$ (5)
A	5	33.375	49.68	12.53	5,112
	6	33.375	41.08	11.28	4,138
	7	33.375	42.41	9.99	3,227
	8	33.375	53.25	10.59	3,639
B	5	33.375	33.92	12.47	4,991
	6	33.375	29.22	11.02	3,902
	7	33.375	32.50	10.00	3,239
	8	33.375	40.53	9.84	3,131

Step 4. Calculate the different kinds of mean scores. At least 6 decimal places should be carried out, if possible. The different kinds of mean scores are given in Table 83.

TABLE 83
DIFFERENT KINDS OF MEAN SCORES

$i \backslash s$	1	2	3	4	$\bar{X}_{.i}$
1	49.68	41.08	42.41	53.25	46.605
2	33.92	29.22	32.50	40.53	34.0425
$\bar{X}_{s.}$	41.800	35.150	37.455	46.890	40.32375 = $\bar{X}_{..}$

Step 5. Calculate the following values:

$$a = N\bar{X}^2_{..} = 267(40.32375)^2 = 434{,}143, \text{ where } N = 8n$$

$$c = 2n \sum_s \bar{X}^2_{s.} = 66.75[(41.800)^2 + \cdots + (46.890)^2] = 439{,}503$$

$$d = 4n \sum_i \bar{X}^2_{.i} = 133.5[(46.605)^2 + (34.0425)^2] = 444{,}678$$

$$e = n \sum_s \sum_i \bar{X}^2_{si} = 33.375[(49.68)^2 + \cdots + (40.53)^2] = 450{,}324$$

Step 6. Calculate the sum of squares for the different sources of variation:

(1) Within subclasses: $\sum_s \sum_i \sum_t x^2_{sit} = (5112 + \cdots + 3131) = 31{,}379$ [Refer to Table 82, column (5)]

(2) Interactions: $n \sum_s \sum_i \bar{X}^2_{si} - 2n_s\bar{X}^2_{s.} - 4n \sum_i \bar{X}^2_{.i} + N\bar{X}^2_{..} = e - c - d + a = 286$

(3) Between grades: $2n \sum_s \bar{X}^2_{s.} - N\bar{X}^2_{..} = c - a = 5360$

(4) Between schools: $4n \sum_i \bar{X}^2_{.i} - N\bar{X}^2_{..} = d - a = 10{,}535$

(5) Total (1) + (2) + (3) + (4) = 47,560

Step 7. Analysis of variance to test different hypotheses. First, we wish to test the hypothesis $H_1: \bar{X}_{11} - \bar{X}_{12} = \bar{X}_{21} - \bar{X}_{22} = \bar{X}_{31} - \bar{X}_{32} = \bar{X}_{41} - \bar{X}_{42}$; or that there is no interaction between grade and school. The results are summarized in Table 84. It is noted that if we have p grades and q schools, then the degrees of freedom for each source of variation are as follows:

Within subclasses	$N - pq$
Interaction	$(p - 1)(q - 1)$
Between grades	$p - 1$
Between schools	$q - 1$
Total	$N - 1$

The additive property of degrees of freedom is clearly demonstrated. From the results in Table 84, we may accept the hypothesis that the interaction is not significantly different from zero. Therefore, we may pool the sum of squares due to "interaction" with "within" sum of squares, as well as the degrees of freedom. We may call this sum "residual"; it can be used as the basis of testing the other hypothesis. (*Note:* If the interaction is significant, we do not pool it with "within.") Next, we wish to test the other two hypotheses, namely, $H_0': \bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \bar{X}_4$ and $H_0'': \bar{X}_{.1} = \bar{X}_{.2}$. The first hypothesis is that there is no difference between the four grade means. The second hypothesis is that there is no difference between the two school means. The results are summarized in Table 85.

TABLE 84
RESULTS OF TESTING THE HYPOTHESIS H_1

Source of variation	D.F.	S.S.	M.S.	Hypothesis
Within subclasses	259	31,379	121.15	....
Interaction	3	286	95.33	Accepted

TABLE 85

ANALYSIS OF VARIANCE FOR SPEED SCORE IN GATES READING SURVEY

Source of variation	D.F.	S.S.	M.S.	F	Hypothesis
Residual	262	31,665	120.86	.....	....
Between grades	3	5,360	1,786.66	14.78	Rejected
Between schools	1	10,535	10,535.00	87.17	Rejected
Total	266				

From results in Table 85, we reject both the hypotheses H_0' and H_0''. Therefore, we conclude that there are significant differences between the means of the grades and that there is also a significant difference between the means of the schools.

PROBLEMS

1. Is there a significant difference among the means of reaction times for age and for sex?

REACTION TIMES IN SECONDS TO LIGHT AND SOUND OF VARIOUS AGE GROUPS (4–60 YEARS) ACCORDING TO SEX

Age group	Male					Female				
	N	Light		Sound		*N*	Light		Sound	
		Mean	S.D.*	Mean	S.D.		Mean	S.D.*	Mean	S.D.
A	10	.34	.1070	.34	.0928	10	.62	.1644	.59	.1890
B	10	.24	.0400	.23	.0409	10	.32	.0340	.31	.0407
C	10	.22	.0331	.19	.0338	10	.26	.0192	.20	.0736
D	10	.26	.0465	.24	.0141	10	.34	.0378	.30	.1139
E	10	.27	.0266	.25	.0467	10	.36	.0342	.30	.0372
F	10	.38	.0574	.37	.0806	10	.44	.0721	.42	.0842

* Standard deviation, Pearsonian.

2. Give a complete analysis of variance for the following data:

REPORTED TESTS WITH STANFORD ACHIEVEMENT TEST BATTERY IN 1924 (DATA FROM BALDWIN)

	Age	Number of cases	Mean	Unbiased S.D.
Boys	9	100	27.4	10.18
	10	117	37.9	11.63
	11	96	44.2	12.95
Girls	9	115	29.1	10.76
	10	126	38.3	10.52
	11	87	44.2	11.04

3. Test the significance of the difference between the means of students in arithmetic computation in the different types of schools, Grade 4.

Arithmetic Computation Scores by Type of School
(After Peterson, 1948)

Score interval	Frequency					Total
	Boarding	Day	Mission	Non-res.	Public	
55–59	0	1	0	0	0	1
50–54	1	1	0	0	0	2
45–49	4	3	0	1	1	9
40–44	4	10	1	0	15	30
35–39	41	60	29	8	90	228
30–34	84	146	48	17	231	526
25–29	80	148	31	17	222	498
20–24	69	166	30	16	129	410
15–19	75	165	24	12	123	399
10–14	48	130	11	8	62	259
5– 9	37	98	8	6	47	196
0– 4	11	36	3	5	19	74
Total	454	964	185	90	939	2632

4. The data on the following page were obtained from the administration of two tests to a random sample of 132 students in a class in college biology. Test 1 was designed to measure the acquisition of fundamental facts and principles; Test 2, to measure the ability to apply a knowledge of facts and principles.

Problem: Test the linearity of regression of scores in Test 2 on scores in Test 1.

Student No.	Score on		Student No.	Score on		Student No.	Score on	
	Test 1	Test 2		Test 1	Test 2		Test 1	Test 2
1	63	34	45	63	34	89	53	27
2	71	42	46	83	44	90	49	22
3	70	41	47	80	52	91	90	49
4	119	50	48	89	49	92	69	31
5	109	57	49	98	44	93	52	37
6	75	30	50	73	35	94	40	41
7	88	33	51	65	30	95	82	40
8	83	55	52	62	30	96	90	37
9	68	20	53	114	54	97	108	54
10	59	35	54	105	39	98	83	40
11	55	43	55	88	35	99	98	37
12	106	47	56	78	49	100	61	18
13	56	35	57	69	51	101	80	39
14	81	51	58	67	36	102	70	40
15	102	48	59	79	29	103	60	30
16	94	43	60	80	38	104	66	34
17	97	40	61	47	36	105	71	31
18	84	39	62	68	42	106	85	46
19	91	51	63	93	44	107	43	26
20	85	41	64	78	37	108	65	32
21	106	49	65	51	34	109	53	35
22	86	49	66	92	46	110	88	45
23	104	41	67	76	36	111	68	41
24	78	40	68	105	57	112	93	46
25	91	51	69	55	32	113	91	47
26	82	43	70	86	50	114	101	56
27	64	34	71	71	30	115	94	40
28	55	38	72	70	31	116	91	41
29	87	40	73	68	28	117	73	33
30	50	30	74	81	39	118	99	47
31	75	46	75	81	48	119	99	45
32	73	41	76	65	39	120	66	40
33	59	43	77	104	49	121	78	40
34	91	48	78	88	43	122	56	37
35	80	52	79	78	32	123	93	48
36	105	59	80	84	40	124	85	38
37	97	48	81	92	47	125	58	36
38	77	39	82	84	35	126	92	43
39	124	52	83	78	48	127	75	31
40	68	34	84	66	25	128	66	27
41	101	49	85	94	53	129	69	44
42	81	34	86	52	39	130	111	50
43	69	44	87	61	38	131	73	35
44	73	40	88	96	43	132	73	41

5. Analyze the following data obtained for Indian students in the twelfth grade showing scores on an arithmetic test and the number of schools attended (after Peterson, 1948):

Arithmetic comp. score	Number of schools attended				
	1	2	3	4	Over 4
95–100	0	1	0	0	0
90– 94	0	0	0	0	0
85– 89	0	0	0	0	0
80– 84	0	0	0	0	0
75– 79	2	2	2	4	0
70– 74	18	16	13	16	9
65– 69	12	39	20	21	20
60– 64	15	56	30	21	8
55– 59	14	48	23	18	18
50– 54	8	46	22	13	13
45– 49	2	30	21	11	8
40– 44	2	16	19	13	6
35– 39	3	17	9	5	3
30– 34	1	5	7	7	0
25– 29	0	7	3	0	0
20– 24	0	3	1	0	0
15– 19	1	2	1	0	0
10– 14	0	0	0	0	0
5– 9	0	1	1	1	0
0– 4	0	1	0	0	0
Total	78	290	172	130	85

6. Test the significance of the difference between the means on the achievement test of the experimental and control groups after adjustment has been made for any inequalities in the two groups with respect to pretest and I.Q. scores. The data on the following pages derive from an experiment to evaluate the effectiveness of the school excursion in teaching a unit on Communication in the sixth grade in eight elementary schools (Clark, 1938).

Primary Data for Schools Comprising Control Groups

Individual	C.A.	M.A.	Pretest score	Final test score
1	153	117	23	31
2	146	146	34	50
3	148	129	40	44
4	142	142	41	51
5	152	137	32	39
6	143	138	40	53
7	141	140	29	38
8	157	145	37	39
9	139	142	32	47
10	141	152	32	49
11	145	158	47	57
12	143	144	38	47
13	146	158	39	50
14	144	130	28	44
15	143	169	36	49
16	147	155	39	48
17	143	158	51	58
18	148	124	15	26
19	151	149	46	53
20	138	172	51	56
21	140	147	34	50
22	143	146	35	40
23	146	140	29	37
24	161	147	24	31
25	142	156	39	46
26	145	171	44	58
27	147	141	32	49
28	134	145	34	52
29	151	141	33	45
30	146	148	39	43
31	153	161	29	42
32	146	141	29	42
33	175	142	28	48
34	144	145	34	39
35	149	144	32	51
36	150	127	24	33
37	144	125	22	29
38	158	134	24	40
39	134	157	55	56
40	142	149	27	37
41	140	140	41	52
42	145	146	43	46
43	146	134	37	46
44	145	152	38	45
45	144	150	54	59
46	157	147	35	45
47	154	145	36	52
48	139	125	16	33
49	145	135	29	38
50	143	150	37	48

PRIMARY DATA FOR SCHOOLS COMPRISING CONTROL GROUPS (*Continued*)

Individual	C.A.	M.A.	Pretest score	Final test score
51	171	124	11	25
52	151	154	35	39
53	163	149	36	44
54	146	148	25	39
55	139	155	35	50
56	149	137	27	35
57	142	128	14	25
58	90	147	33	36
59	140	154	27	38
60	141	143	28	40
61	146	128	8	23
62	146	146	40	56
63	148	152	49	59
64	142	136	29	44
65	160	140	35	40
66	142	154	23	40
67	145	157	41	55
68	146	159	43	57
69	143	136	36	48
70	146	139	40	37
71	143	145	21	38
72	144	134	40	47
73	145	161	37	45
74	146	146	27	34
75	145	148	38	51
76	155	160	29	42
77	143	145	32	49
78	156	126	17	31
79	139	142	23	34
80	146	152	30	43
81	141	166	26	38
82	134	155	33	44
83	142	146	34	39
84	137	151	35	46
85	140	142	26	42
86	138	161	37	48
87	150	142	28	38
88	143	155	38	41
89	143	137	19	34
90	143	151	27	40
91	143	150	25	40
92	143	146	42	51
93	162	131	34	51
94	154	143	43	57
95	158	132	38	48
96	144	149	42	55
97	148	138	35	45
98	145	149	49	67
99	145	147	40	52
100	141	167	56	62

PRIMARY DATA FOR SCHOOLS COMPRISING CONTROL GROUPS (*Continued*)

Individual	C.A.	M.A.	Pretest score	Final test score
101	142	146	32	49
102	141	144	49	62
103	145	141	37	54
104	144	141	46	61
105	139	132	28	35
106	143	139	38	56
107	143	150	40	53
108	143	149	36	50
109	142	135	26	33
110	144	145	24	43
111	148	157	45	52
112	139	147	42	52
113	151	124	35	47
114	141	129	49	50
115	150	134	38	39
116	145	142	42	53
117	147	141	41	56
118	142	142	38	47
119	162	151	34	48
120	184	134	23	44
121	151	126	51	64
122	140	138	37	47
123	141	141	33	44
124	148	134	22	32
125	145	164	47	58
126	164	126	26	38
127	141	147	42	58
128	144	152	36	34
129	149	137	24	42
130	140	157	47	61
131	133	121	38	52

PRIMARY DATA FOR SCHOOLS COMPRISING EXPERIMENTAL GROUPS

Individual	C.A.	M.A.	Pretest score	Final test score
1	145	145	29	48
2	155	154	41	51
3	137	159	47	62
4	138	148	41	54
5	142	156	44	62
6	148	169	41	57
7	148	163	49	62
8	144	146	52	65
9	147	150	47	59
10	145	149	40	54
11	146	137	41	51
12	140	142	42	55
13	147	146	41	51
14	146	153	44	57
15	143	154	27	42
16	140	153	29	44
17	142	140	29	42
18	142	156	39	49
19	139	152	41	56
20	142	149	37	50
21	141	133	24	39
22	138	151	35	51
23	144	142	26	45
24	134	151	38	50
25	142	154	43	58
26	143	138	28	50
27	141	144	35	55
28	141	151	32	53
29	146	153	32	42
30	137	150	47	57
31	135	158	38	52
32	137	163	44	53
33	137	160	45	60
34	148	143	28	45
35	150	142	38	49
36	140	156	52	63
37	127	174	45	59
38	141	143	36	57
39	143	155	41	51
40	139	159	45	58
41	148	142	35	49
42	146	137	39	52
43	145	146	39	50
44	138	146	44	57
45	140	140	36	53

PRIMARY DATA FOR SCHOOLS COMPRISING EXPERIMENTAL GROUPS (*Continued*)

Individual	C.A.	M.A.	Pretest score	Final test score
46	141	149	36	48
47	153	150	27	46
48	140	156	33	40
49	145	166	44	63
50	137	169	20	45
51	146	159	38	52
52	145	130	30	44
53	140	159	32	45
54	141	155	43	57
55	156	140	31	52
56	136	149	37	58
57	140	152	33	57
58	143	138	30	44
59	145	145	32	46
60	145	140	38	55
61	140	160	50	68
62	146	122	23	41
63	140	147	36	50
64	139	162	47	59
65	147	143	37	52
66	143	147	42	61
67	141	137	34	46
68	145	143	36	49
69	137	142	34	50
70	157	120	17	31
71	139	152	41	48
72	146	141	25	43
73	146	137	18	29
74	139	164	39	55
75	129	136	36	46
76	145	163	40	52
77	139	151	15	26

7. Analyze the data in Problem 6, using the Johnson-Neyman technique and setting up the region of significance if it exists. Contrast this technique with that of analysis of variance and covariance (see Ref. 6).

8. How can the analysis of variance technique be used in problems of estimation, that is, in the detection and estimation of components of random variation associated with a composite population? (See Ref. 1.)

References

1. Crump, S. Lee, "The Estimation of Variance Components in Analysis of Variance," *Biometrics*, Vol. 2 (1946), pp. 7–11.
2. Fisher, R. A., *The Design of Experiments*, 2d ed. London: Oliver & Boyd, Ltd., 1937, Section 24.
3. ———, *Statistical Methods for Research Workers*, 10th ed. London: Oliver and Boyd, 1946, Section 39.
4. ———, "The Analysis of Covariance Method for the Relation Between a Part and the Whole," *Biometrics*, Vol. 3 (1947), pp. 65–68.
5. Jackson, Robert W. B., *Application of the Analysis of Variance and Covariance Method to Educational Problems. Dept. of Educational Research, University of Toronto, Bulletin* 11 (1940).
6. Johnson, Palmer O., and Neyman, J., "Tests of Certain Linear Hypotheses and Their Application to Some Educational Problems," University of London, Department of Statistics, *Statistical Research Memoirs* I (1936), pp. 57–93.
7. Newman, Horatio H., Freeman, Frank N., and Holzinger, Karl J., *Twins: A Study of Heredity and Environment.* Chicago: University of Chicago Press, 1937.
8. Snedecor, George W., and Cox, Gertrude M., *Disproportionate Subclass Numbers in Tables of Multiple Classification, Iowa State College Experiment Station Research Bulletin* 180 (1935), pp. 236–272.
9. Tsao, Fei, "Tests of Statistical Hypotheses in the Case of Unequal or Disproportionate Numbers of Observations in the Subclasses," *Psychometrika*, Vol. 7 (1942), pp. 195–212.
10. ———, "General Solution of the Analysis of Variance and Covariance in the Case of Unequal or Disproportionate Numbers of Observations in the Subclasses," *Psychometrika*, Vol. 11 (1946), pp. 107–128.

CHAPTER XII

THE PRINCIPLES OF EXPERIMENTATION

There is an increasingly general realization that a formal experiment is an exacting enterprise designed and carried through with meticulous care to answer a few definite questions. The ability to formulate productive hypotheses and to design experiments to test them is the mark of a first-rate research worker or scientist. An understanding of the principles underlying modern designs is essential at every stage of an experiment if the primary data are to be collected in such a way as to provide the basis for valid inference and so as to enable the maximum amount of information to be elicited from them most efficiently. Perhaps a clearer grasp of the requirements underlying sound experimentation can be gained by the scientific reader through studying and examining designs that lead to valid conclusions. He should apply the techniques to actual problems, however, since difficulties usually tend to disappear on such closer experience.

The whole subject of complex experiments is undergoing rapid development as new possibilities of the methods and of their correct application become better understood. The principles of experimentation, which originated in agriculture, are finding increasing application in many fields of science. The difficulties met with in application in one field are not identical with those in other fields, but many are similar. The solutions of problems arrived at in one field are often of material help in another. Where fields differ fundamentally, new techniques are necessary. Such needs are discovered only in direct contact with the obstacles themselves. Because modifications and extensions of the principles of design are capable of, and will undoubtedly have, ever wider application, the student of modern methods and statistical analysis needs to know how to apply these principles and how to read intelligently the reports of research workers who have used them.

Modern ideas of experimental design differ sharply from earlier or traditional ones. It has long been an admonition in philosophical treatises of scientific experiment to hold constant all except one of the factors in a complex so that its effect may be determined. The experimenter is advised to arrange an experiment so as to make it as sensitive as possible with respect to one question but as insensitive as possible with respect to all others. Just as mathematical development has been biased toward physics, so has the direction of experimentation been largely determined by the pattern of experimentation set up in physics, which emphasizes the importance of varying the essential conditions only one at a time. The difficulty in applying such a principle, particularly in branches of science where the data of the research worker are subject to all sorts of fluctuations, had long been recognized by critical workers. The liberation of the research worker from stereotyped experimentation is relatively recent.

The problem underlying the development of procedures appropriate to deal with types of variable material is twofold; one aspect dealing with the design or logical structure of the experiment, the other with the analysis and interpretation of the results. The development of the logical structure underlying the whole technique of modern experimental design and of the appropriate statistical tools for the analysis and interpretation of the results of such experiments is largely due to R. A. Fisher. Beginning his work in 1919 with the founding of the statistical laboratory at Rothamsted (Harpenden, England), Professor Fisher has revolutionized the science of statistics and the principles of designing biological experiments. His principles of experimentation and methods of statistical analysis are finding increasing application in many fields of science, particularly wherever the basic materials are variable. The possibilities of applying these principles also to the improvement of physical and chemical experimentation have barely been recognized. In biophysics and biochemistry these principles are likely to become increasingly important.

The subject of the design of experiments is too large and too important to scientific workers for it to receive incidental treatment only. In his text *The Design of Experiments* Fisher presents the framework of scientific inference and the principles of modern experimentation. Our discussion is limited to a brief consideration of the major characteristics of modern experimental designs. We are especially interested in the role which statistical procedures play in serving the requirements of sound experimental design and in furnishing the means for unambiguous interpretation.

The Self-contained Experiment. A principle of general utility in statistical analysis is to rely upon the evidence from the data themselves when allowances are to be made for certain inequalities, as in certain comparisons under consideration. Arbitrary corrections based on an a priori basis without reference to the information provided by the data themselves cannot lead to convincing conclusions. Violations of statistical principles of this kind, though not so obvious a misuse of statistical analysis as is an arbitrary selection among observational data previous or subsequent to collection, are probably the source of the political principle that "anything can be proved by statistics," or of the crescendo "lies, damned lies, statistics."

Fisher sets up the self-contained experiment as the model for the research worker and describes the properties which such a model must possess. Although progress in science may result from the better ordering of the experiences we have had, it is chiefly in the collection of new experiences that advancement takes place. However, if these experiences are to afford a secure basis for bringing new knowledge into being, they must be planned in advance in accordance with principles that make such outcomes possible. Thus, experimental observations are essentially experiences formulated at the time of arranging for their collection. Experimental observations are related to existing bodies of scientific knowledge as new observations are carried out to test theories growing out of the previous collection of data. Theories in turn become modified and reformulated as an outcome of the new observations. But once an experiment has been designed and executed, its interpretation must be based on its own evidence. The purpose, therefore, of making an experiment self-contained is to make possible the valid and unequivocal interpretation of its results without referring for decision or settlement or consideration to other experiments or to the aggregate of experiences of prior collection. The principle that an experiment should be self-contained determines the essential difference between mere statistical observations and those which are collected in accordance with a clearly conceived plan.

The Function of Controls. A primary requisite of the principle that an experiment should be self-contained is the necessity of supplying a control or controls, that is, the need to base all conclusions concerning the differential effect of two or more contrasting treatments on the differences in the response or reaction of two or more similar bodies of experimental material. By the use of controls, experiments become comparative and not merely absolute. Absolute information is usually of little interest or importance. The reasoned explanation of the function of controls is clearly illustrated by the following example (Ref. 2).

Assume that an experimenter working with animals injected some fluid into 3 rabbits and found that all 3 got violent and prolonged convulsions followed by death within an interval of 24 hours. In support of his conclusion that the injected substance was the cause of the death of the animals, the experimenter might draw from his own previous experiences or from those of rabbit breeders in general. Admittedly, only rarely would three designated animals die in the way described within such a short period of time. How would the conclusion have been made stronger if the experimenter had taken the precaution to inject a number of control rabbits with a neutral substance at the same time at which he injected his experimental animals? The answer to this question provides the rationale underlying the use of controls. It is that the controls are used to exclude, at a designated level of probability, a number of alternative interpretations of the experimental results—possibilities which have individually and collectively an unknown probability of having occurred. For example, the rabbits might have been ill from tetanus, hydrophobia, cholera, or some other unsuspected epidemic disease; perhaps the needle was infected with a poisonous substance; or it might be that the experimenter's stock was genetically of a kind which reacted in this way in general to injections. Suppose, however, that the experimental rabbits had been randomly chosen from the whole herd, the controls included. Then, if their reaction was clearly different from that of the controls, there was available a precise measure of probability for causes other than the experimental factor for having brought about the observed result. The probability is based exclusively on the number of rabbits used, completely independent of all prior experience of these animals. Assume, for instance, that 5 control rabbits have been selected at random from the total number and, after having been injected with distilled water, had not died of convulsions. The measure of probability is obtainable from a simple application of permutations and combinations.

There are 56 ways of choosing a group of 3 objects out of 8. If the 3 objects were to be selected consecutively, there would be successively 8, 7, and 6 objects to choose from and, therefore, the succession of choices could be made 8 × 7 × 6, or 336, ways. This number represents not only every possible set of 3 but also every possible set in every possible order. Three objects can be arranged in order in 3 × 2 × 1, or 6, ways. The number of possible choices is found by dividing 336 by 6, which is 56. The result, 56, is essential for the interpretation of the experimental results. The 56 sets of 3 which might be chosen would be distributed among the possible events as follows:

Number Dying	*f*
0	10
1	30
2	15
3	1
Total......	56

The probability of the observed difference, if it were not attributable to the material injected, is, therefore, 1 in 56, or a probability level of .018, which by the usual standards may be regarded as significant. It is also worth noting that the use of the controls serves to transform the quality of the experimental evidence by making it strictly objective for others who have not undergone the experiences of the experimenter.

The weight of previous or outside evidence is even much less when the object of the experiment is quantitative, because such evidence is usually very indefinite or highly variable. Thus, the essential condition for controlling the interpretation of experimental results is the provision of comparisons between two or more unlike variants.

The Valid Estimate of Experimental Errors. The second requirement of a self-contained experiment is that it must hold within itself the possibility of securing a valid estimate of the experimental errors which really influence the comparisons made. That is, it is necessary to estimate the error from the data of the experiment itself, because it is only under such conditions that proper confidence can be put in the result of the experiment. In any experiment there are factors which are susceptible to some degree of control by the experimenter. But their effect cannot be entirely eliminated, owing to chance fluctuations. Many of the factors giving rise to these fluctuations which affect performance are small in size and random in incidence, so that it is impossible to present an exhaustive list of all the sources of variation in the experimental material. It is customary to designate the component of variation associated with the random variation of the experimental material as experimental error. The errors do not follow any known exact laws, and so the laws of chance are usually designated as descriptive of their distribution.

As was pointed out in the discussion of analysis of variance, it is assumed that the experimental errors to which the experimental observations are subject shall be independently and normally distributed with the same variance. The importance of the experiment making possible a valid estimate of the experimental errors is indicated by the fact that only under such conditions is it possible to apply to the experimental results tests of their significance which are disconnected from all past experience and are hence capable of adding new knowledge. Therefore, the design of a self-contained experiment involves the consideration of means of affording a valid estimate of error as well as ways of making possible an unbiased comparison between contrasted treatments. The validity of other estimates of error would depend on other mathematical assumptions which the particular method of estimation would introduce. There would be no objective reason for accepting such assumptions as true, if the experimenter has not taken the precautions needed to make them true.

Replication. The first requirement of an experiment designed so that a valid test of significance may be applied in its interpretation is *replication,* the process of repeating the same treatment on more than one object of the experimental test. The word "plot" is used in agricultural experimentation to indicate an individual plot or area of land. The "plot" could be an experimental animal or an individual, for instance. Replication is essential in the first place since it is a means of diminishing the experimental error. Just how this is done may become clear by considering, first, certain factors contributing to the actual errors of the experiment. The amount of information ~~which a particular~~ experiment affords is known as its *precision.* Fisher succeeded in quantifying the concept of information so that now the precision is wholly a quantitative factor in the value of an experiment.

There are a number of factors, both quantitative and qualitative, which may contribute to make the actual errors small. Some of these are the measurement of the criterion; the improvement in the techniques of controlling nonexperimental factors; care in ensuring that in the experimental material the general conditions are those occurring in population practice; the measurement of controls under as nearly as possible the same conditions as those for the unknowns, including time; and the greatest possible avoidance of hidden systematic errors as well as subjective errors. Only when sufficient care has been given to ensure that working errors have been reduced to unimportant quantities can improvement of the replication and the organization of the structure or arrangement of the experiment be expected to achieve greatly increased precision or sensitiveness. The process of reducing working errors begins with reducing the largest sources of error, and it continues until sources of error that hitherto seemed inconsequential become significant by limiting the value of the whole enterprise.

The second function of replication in an experiment is to provide the data from which the appropriate estimate of experimental error can be calculated. Thus replication performs the double service of reducing experimental error and of furnishing an estimate of the error that remains. Replication is the sole source of the estimate of error. To make certain that the estimate of error is unbiased requires as much attention in the design of an experiment as does the guarantee that any of the direct estimates are without bias. Furthermore, the unbiased estimate of experimental error is fundamental for the application of valid tests of significance by which the value and significance of the experiment are determined. Likewise, an unbiased estimate of error is a necessary condition if one is to assess the weight that may be given to the evidence of an experiment should its results differ from those of other experiments of the same sort.

Since the accuracy of an experiment as represented by the standard error of a mean of any one treatment increases in proportion to the square root of the number of replications, it is clearly indicated that a larger difference in treatments would be necessary to demonstrate the significant effect based on a smaller than on a larger number of replications.

The argument is sometimes advanced that the results are good enough if there is reason to believe that the estimate of error is at least not an underestimate. Fisher points out that the danger of the fallacy of assuming to be "on the safe side" is that there is no security in admitting a bias in either direction. The effect of overestimating the error may be to prevent the experimenter from drawing a conclusion which the experiment justly substantiates. Such a practice could lead to the belief that an effect is consequential when it is not, and so to ignore the real cause of disturbance in the design of subsequent experiments. Because of the exploratory and tentative character of much research, a promising line of inquiry might be given up through a failure to discern the clue which the experiment might otherwise have provided.

Randomization. It is essential in an experiment to recognize that equalization is approximate to a greater or lesser degree, no matter how much care and experimental skill are exerted in attempting to equalize the nonexperimental conditions which are likely to influence the result. In many significant practical situations the attempts at equalization are definitely inadequate. It becomes of fundamental importance that this inequality shall not lead to biased estimates and invalid tests of significance. The essential safeguard is included in the experimental procedure by a process which is known as *randomization.* Just how this operation works can be explained by considering again the origin of error. The real errors of the experimental results originate from differences in the nonequalization of the nonexperimental conditions among the objects or groups of objects that are treated differently. The estimates of error are secured from the discrepancies among the objects treated alike. Consequently, it is necessary only to make certain that any two objects that may be treated alike have the same probability of being so treated. Likewise, if treated differently, the objects must have the same probability of being so treated, in each of the ways in which this is possible. This precaution is necessary to assure that each component of error which may influence the experimental results may with equal frequency furnish the data used in the estimate of error. The calculus of probability and the mechanism of the statistical theory of sampling distributions can then be applied with confidence.

Randomization, then, is the procedure of making certain that the probabilities of being subjected to like treatment are equal for every relevant pair of objects in the experiment. It is worthy of note that the object of randomization is not to increase the precision of the experiment but only to guarantee that whatever precision the experimental arrangement is capable of providing is neither over- nor underestimated. Systematic arrangements of plots or objects in contrast to random arrangement have been shown to give consistently either an over- or an underestimate of error.

Controls, replication, and randomization have been discussed as the essential aspects of the principle that an experiment should be self-contained.

Relationship between Experimental Design and Statistical Analysis. The relation between experimental procedure and statistical analysis will now be considered more fully. It is apparent from the discussions of experimental design that a substantial number of the ideas or concepts are of a statistical nature. In fact, a clear understanding of the statistical procedures used is an essential part of the understanding of the principles of experimentation. These procedures serve to fulfill the requirements of intelligible and accurate experimental design and to provide the machinery of unequivocal interpretation. We note, then, that the question of experimental procedure and that of statistical analysis are two aspects of the single problem—the problem of fulfilling the requisites of the operations involved in making additions to scientific knowledge by experimentation.

An analysis of the relationship between the two aspects reveals that once the practical experimental procedure is established, only one method of statistical analysis can be valid. Furthermore, a fact of great practical significance is that the validity of the statistical analysis depends upon the introduction of a random element in the arrangement of the objects of the experiment. A definite and complete statement of this specific process of randomization followed determines in advance the correct statistical method to be applied to the experimental results. The logical organization of each of the possible types of randomization is set forth by the analysis of variance. The neatness of the arrangement of calculations and of the facility of their interpretation in the analysis-of-variance table is greatly appreciated by the modern research worker. The compactness and simplicity of this form of summarizing the results as well as the logical structure of the experiment have added greatly to the intelligibility and accuracy of its interpretation. The logical structure of the experiment is shown by the division of the total number of degrees of freedom, the independent comparisons, corresponding to each of the sum of squares calculated.

The development of principles improving the art of experimentation has been concomitant with that resulting in tools suitable to analysis of experimental results. The standardized methods of statistical analysis were designed largely on the basis of a mathematical theory in which the problems underlying experimental designs of more recent origin had not been explicitly considered. It has been previously pointed out how "Student's" discovery of the *t*-distribution and Fisher's extension to the *z*-distribution made exact tests of significance possible, both for small and for large samples. The modern advances in experimental design have brought about an increased awareness in practical work of the numerous different sources of variation affecting experimental and observational material. Exact tests of significance and the technique of the analysis of variance are indispensable in the assessment of these various components of variation.

We should not overlook the mathematical framework upon which the modern tools of scientific value have been built. This framework gives precision to tests of hypotheses concerning factors giving rise to variation and to experiments planned to yield maximum information.

The statistical treatment of the results of replicated experiments is usually established on the assumption of the normal law of error, and the general formulation of the analysis is drawn from the method of least squares. It is essential for the correct application of the method of least squares that any components of variation not removed by the experimental design be normally and independently distributed. If these conditions are not fulfilled, the theoretical basis underlying tests of significance breaks down and hence estimates and tests of significance are invalidated. Thus, in the test of significance associated with the analysis of variance, it was assumed that the measured effects of the factors under experiment were statistically independent and normally distributed variates, all with the same variance but with possibly different means. Unless, therefore, the arrangement of experiments is balanced to fulfill the assumptions, the statistical reduction of the data would be very difficult, and convincing results would be impossible. Such a balanced arrangement is illustrated in Equation (10.04), page 214, where the entire calculation is much simplified by the fact that when the equation is squared and the terms are summed, the cross-products become zero. Another significant property is that the difference between the means for any one factor is independent of the other factors.

The validity of the method of least squares as the basis for the testing of hypotheses by experimental results was secured by Fisher through the introduction of randomization into the design. It has been pointed out that systematic arrangements are apt to lead to biased results, because the necessary element of randomization is lacking and hence the test of hypotheses through results based on the method of least squares does not produce the same objective validity as does a test on experimental observations obtained from random arrangements.

In spite of the fact that the relation between the material conduct of an experiment and its statistical interpretation must be used in planning conclusive experiments, some experimenters continue to work with variable material without such design and to obtain discordant results incapable of being fitted into a scientific system. Controversies sometimes arise because different experimenters get diverse results for the same problem. In other cases, methods of statistical analysis are employed which result in definitely misleading estimates of error. Also, methods of experimentation are used which cannot give a valid test of experimental results. The common procedure of consulting a statistician or statistical principles after an experiment or investigation has been completed is equivalent to holding a post-mortem analysis. Perhaps the only interpretation of the data that can be made is to state from what the experiment died. But when research workers turn to sound methods of statistical analysis which involve carefully planned experimental designs, difficulties of the type enumerated above tend to disappear.

Therefore, we can state that the most important work of the statistician is to prepare the plan of the experiment or investigation in such a way as to get the best answers to the questions raised. It has been demonstrated that a complete overhauling of the process of collecting, or of the experimental design, can often increase the precision tenfold or twelvefold for the same expenditure in time and labor. The modern research worker, therefore, needs statistical knowledge not only for working out the results but also for designing: unless he has a working knowledge of the technique he employs, he cannot conduct his experiment properly. In planning an experiment, it is especially important to give due attention to possible results and their interpretation. The experimenter must be induced to use his imagination, and to anticipate the confusion and difficulties that will assail his investigation if they are not foreseen.

References

1. Fisher, R. A., *The Design of Experiments,* 4th ed. Edinburgh: Oliver & Boyd, Ltd., 1947.
2. ———, "The Independence of Experimental Evidence in Agricultural Research," *Oxford Papers, Third International Soil Conference,* 1935, pp. 112–119.
3. Wishart, M. A., and Sanders, H. G., *Principles and Practice of Field Experimentation.* London: The Empire Cotton Growing Corporation, 1935.
4. Yates, F., " The Principles of Orthogonality and Confounding in Repiicated Experiments," *Journal of Agricultural Science,* Vol. XXIII (1933), pp. 108–145.

CHAPTER XIII

APPLICATIONS OF THE PRINCIPLES OF EXPERIMENTATION

We now proceed to show the application of the principles of experimentation to certain cases of technical importance. Our emphasis is upon the interpretation of the experimental results and the fundamental part which statistical methods, particularly those of analysis of variance and covariance, play in this process.

Let us take one of the simplest designs planned to compare the actions of two like individuals under contrasting conditions. A biologist might wish to determine the effect of the removal of a deep-seated organ of an animal. As a control he would perform a similar operation upon another animal of the same kind but in which the organ under investigation would not be disturbed. In this way the experimenter attempts to make the situations alike in all respects except the factor to be tested. Such perfect experimental control is an ideal desideratum which is never capable of complete fulfillment. It is, however, a basic principle upon which experimentation depends.

The Single-Factor Experiment. The method of pairing takes into account two desiderata in experimental design: (1) The requirement of homogeneous experimental material so that the sensitivity of each individual observation may be enhanced, and (2) the need for multiplying the number of observations in order to reveal the reliability and the consistency of the results. The two coupled individuals would, presumably, react alike under the same treatment, and the difference observed under contrasting treatment measures the differential treatment effect. A minimum of two pairs, or replications, is required, since with a single pair it would be impossible to ascribe any difference in behavior detected to the difference in treatments or to the particular variability of the individuals, or to both jointly. The differences between the measurements of the respective pair members constitute the experimental data upon which inferences are to be drawn. Which individual of a particular pair shall receive the one or the other of the two treatments is determined by a random process. If treatments are randomly assigned, replication serves to equalize the effect of uncontrolled sources of variation. It is the variation among the several differences that is used in estimating experimental error. By comparing the mean difference attributable to the differential effect of the treatments with the standard error of the mean difference, the significance of the results of the experiment is to be determined.

We have previously examined the statistical method for reducing the data obtained from an experiment purporting to be of a single-factor type (page 75). The difference between the achievement scores of two individuals paired on the basis of their potential learning capacities was computed for each of the 25 pairs. The null hypothesis was tested that these differences constituted a random sampling from a population of such differences distributed about a mean of zero in a normal manner. The criterion, t, was set up for testing the first aspect of this hypothesis.

The method of replicated comparison of individuals, by pitting each individual against another individual of like kind in conditions made as equal as possible, is a simple and effective experimental design for testing the differential effect between two treatments. It is, however, limited to situations where the presumed effect of a single factor can be measured under the controlled conditions prescribed for the validity of the method. In practice, these conditions are not often present. Furthermore, it is usually desirable to test the effects of more than two treatments. The need for broadening the scope and comprehensiveness of experimental inquiries has led, therefore, to the extension of replicated comparisons of individuals or groups of individuals to more and more complex situations. In this extension, the subdivision of the experimental material into relatively homogeneous series is a fundamental part of the process, as was observed in the paired experiment. Just as the advance in systematic sampling has been made possible by utilizing prior knowledge of the population sampled, so the utilization of knowledge of how to subdivide the experimental material profitably has played an important part in the evolution of experimental design. The principle that the process of subdivision can be advantageously duplicated is also operative. The smallness of number or quantity of sufficiently homogeneous material circumscribes the number of different treatments rather than the number of replications that can be incorporated into an experiment.

The Randomized-Block Arrangement. The experimental design known as the *randomized block* is a simple application of an experimental arrangement illustrating the principle of the subdivision of the experimental material into relatively homogeneous series. In this arrangement each treatment occurs equally frequently, more commonly once in each block, and the treatments are randomly allotted to the experimental units within the block. The term "block" may denote any group containing the required number of experimental units. In arranging the grouping so that similar experimental units are contained in the same block, the accuracy of the treatment comparisons is increased by eliminating from them the differences due to dissimilarities among the different blocks. The process of randomization guarantees that no treatment bias is introduced and permits an unbiased estimate of experimental error basic for the validity of the test of significance.

Consider an experiment in nutrition on the relative effect of 4 different treatments A, B, C, and O (no treatment), which are randomly applied to 4 blocks of 4 children each chosen as nearly alike as possible with respect to age, height, and weight at the beginning of the experiment. The arrangement is represented in the following diagram:

	Block I	Block II	Block III	Block IV
Children Treatment	1 2 3 4 O B C A	5 6 7 8 A C O B	9 10 11 12 B A C O	13 14 15 16 O C B A

We give the analysis for the general case where k denotes the number of blocks and p the number of treatments. Then the equation for the sums of squares is

$$\underset{(1)}{\sum_{1}^{pk}(X-\bar{X})^2} = \underset{(2)}{p\sum_{1}^{k}(\bar{X}_b-\bar{X})^2} + \underset{(3)}{k\sum_{1}^{p}(\bar{X}_t-\bar{X})^2} + \underset{(4)}{\sum_{1}^{pk}(X-\bar{X}_b-\bar{X}_t+\bar{X})^2} \qquad (13.01)$$

where $\bar{X}_b$ is the mean of a block, $\bar{X}_t$ is the mean of a treatment, and $\bar{X}$ is the grand mean. The corresponding equation for the degrees of freedom is

$$\underset{(1)}{pk-1} = \underset{(2)}{(k-1)} + \underset{(3)}{(p-1)} + \underset{(4)}{(p-1)(k-1)} \qquad (13.02)$$

The following formulas are used to calculate the sums of squares:

(1) Total: $$\sum_{1}^{pk}(X-\bar{X})^2 = \sum_{1}^{pk}(X^2) - \frac{T^2}{pk}$$

(where T = grand total for all plots)

(2) Blocks: $$p\sum_{1}(\bar{X}_b-\bar{X})^2 = \sum_{1}^{k}\frac{(T_b^2)}{p} - \frac{T^2}{pk}$$

(where T_b = total for one block)

(3) Treatments: $$k\sum_{1}^{p}(\bar{X}_t-\bar{X})^2 = \sum_{1}^{p}\frac{(T_t^2)}{k} - \frac{T^2}{pk}$$

(where T_t = total for one treatment)

(4) Error: $$\sum_{1}^{pk}(X-\bar{X}_b-\bar{X}_t+\bar{X})^2 = (1) - (2) - (3)$$

(subtract blocks and treatments from total)

These components are then set up in the conventional analysis-of-variance table.

The standard error of the experiment is

$$s = \sqrt{\frac{\sum_{1}^{pk} (X - \bar{X}_b - \bar{X}_t + \bar{X})^2}{(k - 1)(p - 1)}} \tag{13.03}$$

The standard error for the mean of one treatment is

$$s_{\bar{x}_t} = \frac{s}{\sqrt{k}} \tag{13.04}$$

TABLE 86

THE SCORES OF 25 PAIRS OF STUDENTS SUBJECTED TO TWO DIFFERENT TREATMENTS IN A RANDOMIZED-BLOCK ARRANGEMENT

Pairs	Treatments		Difference	Sum
	X_1	X_2	$X_1 - X_2$	$X_1 + X_2$
(1)	(2)	(3)	(4)	(5)
1	73	58	15	131
2	52	37	15	89
3	100	53	47	153
4	60	77	−17	137
5	75	51	24	126
6	67	62	5	129
7	61	55	6	116
8	59	30	29	89
9	33	39	− 6	72
10	19	16	3	35
11	32	15	17	47
12	27	37	−10	64
13	68	44	24	112
14	54	27	27	81
15	26	43	−17	69
16	30	27	3	57
17	69	53	16	122
18	43	29	14	72
19	23	13	10	36
20	11	17	− 6	28
21	26	20	6	46
22	30	9	21	39
23	28	35	− 7	63
24	53	21	32	74
25	23	42	−19	65
Sum	1142	910	232	2052
Sum of squares	64,226	40,474	8962	200,438

We may further illustrate the principles of the randomized-block design by applying them to the experiment presented on page 75. Here there are only two treatments, which are assumed to have been randomly assigned to the members of the respective 25 pairs. Each pair corresponds to a block, and each individual in a pair or block is an experimental unit. The logical structure of this type of experimental design, specified by the process of randomization carried out, is sorted out by the analysis of variance. In this case each item is classified by two criteria; for example, an individual achievement score is classified by treatment and membership in a particular pair. The analysis is carried out as follows:

Step 1. The measures of treatment effects for the respective members of the 25 pairs are given in columns (2) and (3) of Table 86. The difference and the sum of the treatment effects are given in columns (4) and (5). The sum and sums of squares are calculated and recorded in the last two rows, respectively.

Step 2. Calculate the sum of squares for differences:

$$\sum (X_1 - X_2)^2 - \frac{[\Sigma(X_1 - X_2)]^2}{n} = 8962 - \frac{(232)^2}{25} = 6809.04$$

This sum of squares is then divided by 2, the number of achievement scores involved in each difference. This is done to obtain the per-individual measure of the variation of these differences, since the variance of the difference is an estimate of $2\sigma^2$ (see page 37). The quotient of $6809.04/2 = 3404.52$ is entered in Table 87 as interaction or experimental error. If the differences among the individuals of the respective pairs had been the same, there would have been no interaction. Thus, the source of measurement of the experimental error is the uncontrollable variation of these differences.

TABLE 87
ANALYSIS OF VARIANCE OF THE ACHIEVEMENT TEST SCORES IN ALGEBRA OF THE 25 PAIRS OF STUDENTS

Source of variation	D.F.	Sum of squares	Mean square	*F*	Hypothesis
Interaction or experimental error	24	3,404.52	141.855		
Between pairs	24	16,004.92	666.870	4.70	Rejected
Between treatments	1	1,076.48	1076.480	7.58	Remains in doubt
Total	49	20,485.92			

Step 3. Compute the sum of squares from the sums:

$$\sum (X_1 + X_2)^2 - \frac{[\Sigma(X_1 + X_2)]^2}{n} = 200,438 - \frac{(2052)^2}{25} = 32,009.84$$

Here, as in the case of the difference, since the sum is made up of two achievement scores, the comparable sum of squares for pair variation is $\frac{1}{2}(32{,}009.83) = 16{,}004.92$. This value is entered in Table 87 as the "between" pairs source of variation.

Step 4. The sum of squares to measure the variation assigned to treatment effects is obtained as follows. The mean of the two treatment totals is

$$\tfrac{1}{2}\Sigma(X_1 + X_2) = \tfrac{1}{2}(2052) = 1026$$

The two deviations are $1142 - 1026 = 116$; and $910 - 1026 = -116$. The sum of their squares is 26,912. This sum is required on a per-pair basis and is therefore divided by 25. The quotient is entered as the measure of variation due to treatment in Table 87.

Step 5. The total sum of squares calculated independently provides a check on the calculations. It is given by

$$\left[\sum (X_1)^2 + \sum (X_2)^2\right] - \frac{\sum (X_1 + X_2)^2}{2n} = 104{,}700 - \frac{(2052)^2}{50}$$
$$= 20{,}485.92$$

This value is recorded in the total row of Table 87.

Step 6. The total number of degrees of freedom is 1 less than the number of individual achievement scores, or $50 - 1 = 49$. The 25 differences and the 25 sums each contribute 24 degrees of freedom; the two treatments, 1. Thus, the additive property applies to the degrees of freedom as well as to the sum of squares.

Step 7. Tests of significance can now be applied to the results recorded in the analysis-of-variance table. The differential effect due to variation in treatment is found to be $F = 1076.48/141.855 = 7.58$, a value significant at the 5 per cent level. The table values for F corresponding to d.f. 1 and 24 are: $F_{.05} = 4.26$; $F_{.01} = 7.82$. A similar finding was given by the t-test (page 78), where $t = 2.75$ and $t_{.01} = 2.797$ for d.f. $= 24$. This is a demonstration of the fact pointed out on page 55, that if there is only 1 degree of freedom as in this experiment of two treatments, $F = t^2$. Thus, $t^2 = 86.1184/11.3484 = 7.58$.

The test of significance for the differences between the means of the pairs is given by $F = 666.87/141.855 = 4.7$. This value is significant at the 1 per cent level; the value of F for d.f.'s of 24 and 24 is $F_{.01} = 2.66$.

The separation of the source of variation among the pairs illustrates the contribution of the experimental design to the precision of the experiment. If this source of variation had not been isolated, the variations among the pairs would have been included in the experimental error, thus substantially reducing the precision (see Table 88). Thus by using the randomized-block design in this case and putting equated individuals in each block, the variation among pairs has been controlled and isolated.

TABLE 88

Analysis of Variance of the Achievement-Test Scores of the 25 Pairs of Students without the Isolation-of-Treatment Effect

Source of variation	D.F.	Sum of squares	Mean square	F	Hypothesis
Between pairs	24	16,004.92	666.87	3.38	Rejected
Within pairs	25	4,481.00	179.24		
Total	49	20,485.92			

An objective basis for determining the increase in precision in using randomized blocks as compared with the use of two groups of random samples of students for the experimental comparison has been given by Yates (Ref. 21). The calculations are as follows:

The error variance, 141.855, is substituted for the mean square of error (24 D.F.) and the mean square for treatment (1 D.F.). The corresponding sum of squares is found by multiplying the error variance by the combined degrees of freedom. Thus, (141.855)(25) = 3546.375. This product is added to the sum of squares for "between," 16,004.92. Thus, 16,004.92 + 3546.375 = 19,551.295. This sum is then divided by the total degrees of freedom, 49. Thus, 19,551.295/49 = 399.005. The efficiency of randomized blocks as compared to random sampling equals 399.005/141.855 = 2.81 or 281 per cent.

Symmetrical Incomplete Randomized-Block Design. A useful modification of the randomized block type of arrangement is the one known as the *symmetrical incomplete randomized-block design.* In this arrangement each block contains two units only, and all possible combinations of the treatments, taken in pairs, are included in the different blocks (Ref. 21). This type of design has proved to be especially valuable in situations where the experimental material is naturally divisible into groups, with members less than the number of treatments all of which might be of experimental interest. The study of several treatment effects on such homogeneous groups as twins or triplets is an example.

The Latin-Square Design. The experimental principle that the process of subdivisions of the experimental material may be advantageously duplicated is best illustrated by the arrangement known as the *Latin square.* This type of design is similar in principle to a randomized-block arrangement, but in a Latin square two cross-groupings of the experimental units are carried out, corresponding to the rows and columns of a square. The treatments are subject to the double restriction that each treatment occurs once and once only in each row and in each column. Thus, the differences between rows and columns can be eliminated from the experimental comparisons.

The appropriate process of randomization, necessary to ensure the validity of the test of significance applied to the experiment, consists in taking any square arrangement which fulfills the conditions of a Latin square and rearranging either the rows or the columns, or both, at random, and then assigning the treatments at random. The special methods which have to be used to assure complete randomization can be carried out by using the typical "transformation sets" tabulated by Fisher and Yates (Ref. 8).

The structure of a Latin-square design is illustrated in Figs. 6 and 7 and the appropriate statistical analysis follows.

		Suit Presented				
		♣	♦	♥	♠	
Suit Guessed	♣	2	1	6	4	13
	♦	5	3	6	3	17
	♥	4	0	4	0	8
	♠	4	1	4	3	12
Total		15	5	20	10	50

Figure 6. Record for a single-individual.

A	B	D	C
D	A	C	B
C	D	B	A
B	C	A	D

Figure 7. A 4 × 4 Latin square.

Consider an experiment designed to test the telepathic powers of a large sample of individuals. Suppose that the experiment consists in presenting 50 playing cards in sequence, each card being drawn at random from the pack and then returned. Each subject reports his guess of the suit of the card drawn each time. Figure 6 is the record of a single individual. His score of correct assignments is the total of the frequencies in the diagonal cells, for example 12. No 2 cells of a set in the contingency table are in the same row or column and no cell is common in 2 sets. The sets may be defined by the letters of a Latin square as in Fig. 7.

More generally, let the letters A, B, C, D represent treatments in the 4 × 4 Latin square. The "plots" are arranged in 4 rows and 4 columns and there must be as many treatments as there are rows and columns. The treatments are randomly assigned to the plots subject to the double restriction that the treatment can occur only once in any row or column.

We give the analysis for the general case where n represents the number of rows, columns, and treatments. The equations for the sums of squares and degrees of freedom are as follows:

$$\left.\begin{aligned}&\underset{(1)}{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2} = \underset{(2)}{n\sum_{r=1}^{n}(\bar{X}_r-\bar{X})^2} + \underset{(3)}{n\sum_{c=1}^{n}(\bar{X}_c-\bar{X})^2}\\ &+\underset{(4)}{n\sum_{t=1}^{n}(\bar{X}_t-\bar{X})^2} + \underset{(5)}{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij}-\bar{X}_r-\bar{X}_c-\bar{X}_t+2\bar{X})^2}\end{aligned}\right] \qquad (13.05)$$

where $\bar{X}_r$ and $\bar{X}_c$ represent the means of rows and columns, respectively; $\bar{X}_t$ is the mean of a treatment; and X_{ij} is the value of the item in the ith row and the jth column.

The corresponding equation for the degrees of freedom is

$$\underset{(1)}{(n^2-1)} = \underset{(2)}{(n-1)} + \underset{(3)}{(n-1)} + \underset{(4)}{(n-1)} + \underset{(5)}{(n-2)(n-1)} \qquad (13.06)$$

The calculations for the sums of squares are as follows:

(1) Totals: $$\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij}-\bar{X})^2 = \sum_{1}^{n}\sum_{1}^{n}(X^2) - \frac{T^2}{n^2}$$

(T = grand total of all plots)

(2) Rows: $$n\sum_{r=1}^{n}(\bar{X}_r-\bar{X})^2 = \sum_{1}^{n}\frac{(T_r^2)}{n} - \frac{T^2}{n^2}$$

(T_r = total for one row)

(3) Columns: $$n\sum_{c=1}^{n}(\bar{X}_c-\bar{X})^2 = \sum_{1}^{n}\frac{(T_c^2)}{n} - \frac{T^2}{n^2}$$

(T_c = total for one column)

(4) Treatments: $$n\sum_{t=1}^{n}(\bar{X}_t-\bar{X})^2 = \sum_{1}^{n}\frac{(T_t^2)}{n} - \frac{T^2}{n^2}$$

(T_t = total for one treatment)

(5) Error: $$\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij}-\bar{X}_r-\bar{X}_c-\bar{X}_t+2\bar{X})^2 = (1)-(2)-(3)-(4)$$

The standard error in a Latin square is

$$s = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij}-\bar{X}_r-\bar{X}_c-\bar{X}_t+2\bar{X})^2}{(n-2)(n-1)}} \qquad (13.07)$$

The standard error for the mean of one treatment is

$$s_{\bar{x}_t} = \frac{s}{\sqrt{n}} \qquad (13.08)$$

ANALYSIS OF VARIANCE OF THE LATIN SQUARE

Source of variation	D.F.	Sums of squares	Mean square	Variance ratio
Rows........	$(n - 1)$	(2)	$\frac{(2)}{n - 1}$	F_1
Columns.....	$(n - 1)$	(3)	$\frac{(3)}{(n - 1)}$	F_2
Treatments..	$(n - 1)$	(4)	$\frac{(4)}{(n - 1)}$	F_3
Error........	$(n - 2)(n - 1)$	(5)	$\frac{(5)}{(n - 2)(n - 1)}$	
Total	$(n^2 - 1)$	(1)	$\frac{(1)}{(n^2 - 1)}$	

A Greco-Latin Square. A Greco-Latin square is formed by a pair of Latin squares—one written with Latin, the other with Greek, letters—which, when superimposed, possess the property that each Latin letter occurs once in each row and in each column, and each Greek letter appears once in each row and in each column, and with each Latin letter (Ref. 7). Thus:

Aα	Bβ	Cγ
Bγ	Cα	Aβ
Cβ	Aγ	Bα

The two squares are orthogonal to each other. *Orthogonality* is that property of an experimental design which makes possible the direct and separate estimates of each of the several effects. From analytical geometry it is recalled, for instance, that two planes,

$$ax + by + cz + d = 0 \qquad \text{and} \qquad a'x + b'y + c'z + d' = 0$$

are orthogonal (perpendicular) if $aa' + bb' + cc' = 0$. The principle of orthogonality is a basic one in modern experimental designs.

A Latin-Square Design in Psychology. Although the Latin square was originally designed in agricultural experimentation to eliminate from the experimental comparisons possible differences in soil fertility among plots in rows and in columns, it has found useful application in other fields. It is especially advantageous when the disturbing effects of two factors need to be eliminated from the experimental comparisons. In experiments in psychology, for example, the effect of the sequence or order of the experimental factors or situations in space or in time may need elimination.

Thus, in an experiment (Ref. 9) the object was to find out the effect upon recognition of colors when they were presented to the dark-adapted eye of the subject under different degrees of illumination. The following analysis-of-variance table reveals the skeleton of the experimental design and the corresponding divisions of sums of squares and mean squares into the several sources of variation:

Source of variation	D.F.	Sum of squares	Mean square
Among orders of presentation (rows).......	3	Ss_0	$\frac{Ss_0}{3}$
Among illumination levels (columns).......	3	Ss_I	$\frac{Ss_I}{3}$
Among colors...........................	3	Ss_c	$\frac{Ss_c}{3}$
Experimental error.....................	6	Ss_e	$\frac{Ss_e}{6}$
Total	15	Ss_T	

$$F = \frac{\text{mean square (color)}}{\text{mean square (error)}}$$

There were 4 colors—yellow, green, blue, and red—and 4 levels of illumination. Each color was presented once in the first, second, third, and fourth order of presentation at each of the four levels of illumination, and once in the first, second, third, and fourth place in each series order. The colors were arranged at random with this double restriction. It is to be noted that, since all treatments are equally represented in all rows and all columns, no part of treatment differences is included in the row and column comparisons. Thus, the effects of order and illumination levels were removed from the measurement of accuracy of recognition of colors. The measurement consisted in the percentage of the experimental subjects who identified each color correctly. In order to apply the analysis of variance to the percentages, a prior transformation of the data was necessary (see page 165).

An extension of the experiment to determine the effect of the form of the stimulus would require the measurement of the combined effect of color and form. For this purpose the Greco-Latin square could be used, in which each color and each form would be combined so that one and only one combination of each color form occurs. Such color-form combinations would then be handled as a Latin square.

Factorial Design. A formal experiment is designed and executed with meticulous care to provide answers to definite questions. The worth of the experiment is contingent on how wisely the questions have been conceived and formulated. It is fundamental to understand thoroughly the purpose and ultimate applicability of the experiment. A big advantage for complex experiments, that is, those designed to secure answers to a number of definite questions, lies in the fact that they afford results of wider applicability than do simple ones. Until recently, it was regarded as essential that an experiment should be simple and restricted to answering a single question regarding the effect of a single factor. It is important in setting forth the plans of an experiment to answer the questions which prompted the research, to list all the variables that might conceivably influence the results. Due attention must be given to the possible results and their interpretation. Even after listing all the variables that occur to the experimenter, there are others which are not suspected. As many as possible of the variables need to be controlled. However, it is usually desired to secure comparisons under a wide range of conditions of certain variables. In carrying out comparisons of two treatments, for instance, under the same conditions, the relative efficacy may be accurately determined under certain fixed conditions. However, unless these experimental conditions duplicate the practical conditions, the findings of the former may not be applicable at all to the latter. An average value of the ratio of the measures of the treatment effects over a range of conditions is usually the quantity wanted in practical application. In experiments based on the assumption of controlling all factors except the one under investigation, it is often observed that the results change from one experiment to another of the same kind. The difficulty or impossibility of controlling or isolating the various factors involved in experimentation precluded conclusive results in most cases of the traditional "controlled" experiment. Furthermore, as pointed out above, it is usually most important to observe the effects of factors in as nearly a natural setting as possible.

The desideratum in experimentation of observing the effects of varying all the essential conditions simultaneously rather than one at a time attains a substantial realization in the modern methods of design devised to cope with this problem. A very considerable advance has been brought about by the factorial design in experimentation. In this design, all the factors to be examined are varied concurrently in all possible combinations. The principal advantages of this type of design over the traditional experiment planned to examine a single question, or a single factor, consist in its greater efficiency and comprehensiveness. This superiority is achieved through the fact that in a factorial experiment, every trial contributes to the answering of every question with almost the same precision as though the whole experiment had been given over to any one of them. In addition to measuring the effect of each of the single factors, the measures of the effects of the interaction of all combinations of factors are made with the same precision. The latter advantage is especially great, since, with separate single-factor experiments, information could not possibly be deduced concerning the interaction of the different factors.

The investigation of the interactions, though a highly important consideration, frequently was overlooked completely until appropriate means for the measurement of these interactions were developed. A third distinct advantage of factorial design is that this plan gives results of wider applicability than do single experiments, since the exact standardization of experimental conditions prescribed for the traditional experimental design gives information only in respect to a narrowly restricted set of conditions. In the factorial design the ingredients may be varied, that is, applied at different levels, whereas in the single-factor experiments standardization requires that the other factors be kept constant. Rarely is it possible to achieve the degree of standardization required for conclusive results.

A Factorial Experiment in Psychology. The principles of factorial design are illustrated by presenting the design and the analysis of the results of an experiment in psychology.

The psychological experiment[1] consisted in determining the difference limen (D.L.) of subjects for weights increasing at constant rates. Seven different standard weights—100, 150, 200, 250, 300, 350, and 400 grams—and four different rates of 50, 100, 150, and 200 grams per 30 seconds were used. Four men and four women constituted the experimental subjects. Two of each sex were normally sighted; two of each sex were congenitally blind. Five difference limen values were determined for each subject on each of the 28 rate-weight combinations. The order of presentation of each combination was established in advance by the use of Fisher and Yates's set of random sample numbers. The reality of the subject's response was checked by catch stimuli randomly introduced. The entire experiment was repeated on each subject after an interval of one week. Thus, there were 280 D.L.-values for each of the eight subjects. The experimental arrangement may be called a $4 \times 7 \times 2 \times 2 \times 2$ factorial design, that is, the combination of 4 rates, 7 weights, 2 sights, 2 sexes, and 2 dates.

The mean D.L.-value of five trials for each individual on each of the weight-rate combinations for each of the 2 dates was the basis of our statistical analysis. Let us designate the notations for the different variables. The individuals were classified into two sexes, the male being denoted by I and the female by II. Each sex was classified into two sights: the normal denoted by A and the congenitally blind by B. Each individual tried seven different weights: the weight of 100 grams is denoted by 1; of 150, by 2; of 200, by 3; of 250, by 4; of 300, by 5; of 350, by 6; and of 400, by 7. Each weight is combined with each of the four rates: 50 grams per 30 seconds is denoted by *a*; 100, by *b*; 150, by *c*;

[1] For a detailed description of the experiment, the mathematical solution of the problem, and the complete analysis and interpretation of the results, see Ref. 14. The assumption underlying the analysis of variance, that experimental errors are normally distributed with a common variance, was studied by plotting the errors both for totals and subgroups. Within the limitations of the method, the assumptions appeared satisfied.

TABLE 89

MEAN D.L. (MEASUREMENT IN GRAMS) OF ALL INDIVIDUALS ON DIFFERENT COMBINATIONS: X_{sijklt}

Sex		I								II							
Sight		A				B				A				B			
Date		α		β		α		β		α		β		α		β	
Weight	Rate Individual	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	*a*	4.5	14.0	3.1	13.4	24.2	19.3	41.2	19.9	18.5	3.1	11.2	3.9	9.6	9.0	6.1	8.6
	b	10.9	30.5	6.2	26.2	48.1	41.0	59.1	44.1	22.3	7.0	21.8	11.4	17.9	16.1	13.9	14.5
	c	15.8	35.0	11.2	34.6	60.9	56.5	75.8	44.7	27.2	12.2	39.4	12.5	26.1	19.8	17.3	21.7
	d	20.2	43.4	13.3	58.2	69.9	78.2	148.3	73.0	33.2	14.5	36.8	21.8	35.7	32.5	27.3	25.7
2	*a*	4.5	13.9	3.3	12.8	25.3	21.1	29.8	18.3	10.2	3.9	8.8	5.1	7.3	6.4	6.0	6.9
	b	10.1	25.5	6.8	23.5	41.2	30.1	59.7	28.5	25.2	5.4	14.0	10.2	18.2	15.9	12.5	14.2
	c	18.6	39.3	11.4	36.9	52.0	52.8	79.9	48.4	41.1	9.6	28.2	14.8	21.3	21.2	18.5	21.8
	d	23.6	43.8	15.8	50.3	76.7	57.7	123.1	57.0	36.0	15.3	30.5	18.1	27.9	24.3	23.4	26.6
3	*a*	4.5	12.2	3.5	12.8	25.1	19.6	28.5	15.9	11.4	3.6	8.1	3.7	6.5	6.9	6.4	5.3
	b	8.6	29.3	7.0	27.8	31.4	35.0	48.7	31.8	19.7	7.3	21.8	8.7	15.8	12.9	12.1	12.0
	c	12.2	47.8	11.7	37.5	58.2	59.4	69.1	36.0	32.2	11.7	29.0	17.0	22.1	24.1	17.7	20.3
	d	22.3	47.6	14.8	51.0	82.4	59.9	73.5	69.1	36.7	15.1	36.3	20.5	26.2	24.7	23.4	27.2
4	*a*	5.6	15.3	3.4	13.1	17.6	17.1	23.8	14.3	11.0	3.3	9.3	4.9	5.8	6.6	5.7	6.2
	b	9.0	30.5	7.4	25.6	30.4	32.2	38.1	22.8	15.4	5.2	17.0	7.7	12.5	13.4	9.8	13.0
	c	12.8	38.2	11.0	36.4	60.6	47.7	64.4	36.3	21.3	12.4	26.5	12.9	18.2	19.7	14.6	18.8
	d	20.0	59.7	15.2	53.3	76.4	65.4	61.9	59.1	31.0	18.9	31.3	17.3	27.6	28.8	23.2	25.8
5	*a*	3.8	13.0	3.4	12.1	20.7	16.7	20.9	13.0	9.6	3.5	8.6	4.8	5.3	7.2	5.1	6.1
	b	10.6	30.4	7.1	26.4	39.9	29.1	30.7	30.9	15.6	5.3	17.2	11.2	12.5	12.0	9.0	12.1
	c	16.5	57.7	11.7	39.8	57.1	42.2	42.2	34.6	33.7	11.9	32.6	14.1	16.1	21.5	17.2	17.7
	d	20.0	52.6	14.0	52.5	71.4	59.5	77.8	51.8	36.0	16.9	33.0	19.1	23.9	30.0	21.0	25.2
6	*a*	5.1	15.0	3.5	12.7	19.4	14.4	17.8	12.2	7.8	3.6	8.2	3.9	5.5	6.5	4.8	6.5
	b	10.7	34.7	8.2	27.7	36.7	27.8	28.4	28.9	16.5	5.5	15.5	10.7	11.8	12.9	10.9	11.4
	c	15.8	39.7	11.3	39.4	57.9	36.6	53.1	39.7	28.2	12.8	24.0	14.8	17.7	19.6	13.5	17.6
	d	20.9	65.2	14.7	59.4	76.9	53.4	56.0	49.8	37.6	16.2	35.0	19.1	22.2	28.4	23.2	23.5
7	*a*	4.4	17.6	3.8	15.4	17.3	13.5	13.4	12.9	10.9	4.2	8.2	4.9	6.4	7.3	5.7	6.2
	b	9.6	32.5	8.0	24.5	35.5	31.9	27.2	24.3	18.4	7.6	20.4	9.2	11.1	13.1	9.2	12.7
	c	17.0	44.8	10.2	44.4	49.5	38.9	36.3	31.6	29.6	10.5	24.4	15.6	16.7	21.3	13.7	18.6
	d	19.9	58.4	16.5	49.1	79.6	52.1	53.2	45.9	32.8	15.9	33.2	18.4	19.7	24.0	17.0	22.1

and 200, by d. Observations of each individual trial were obtained on two different dates: the first date is denoted by α, and the second by β. Hence, we have $2 \times 2 \times 7 \times 4 \times 2 = 224$ subgroups. Furthermore, we have two people for each subgroup, denoted by (1) and (2). Altogether, then, we have 448 measures of D.L.-values (see Table 89).

Mathematically, each measure is denoted by X_{sijklt}, which is the score made by the tth individual of the sth sex and the ith sight for the jth weight and the kth rate on the lth date. The mathematical expression of the D.L.-value of the tth individual in the sth sex of the ith sight on the jth weight of the kth rate at the lth date is

$$\left.\begin{aligned} X_{sijklt} = A &+ B_s + C_i + D_j + E_k + F_l + I_{si} + I_{sj} + I_{sk} \\ &+ I_{sl} + I_{ij} + I_{ik} + I_{il} + I_{jl} + I_{jk} + I_{jl} \\ &+ I_{sij} + I_{sik} + I_{sil} + I_{sjk} + I_{sjl} + I_{skl} \\ &+ I_{ijk} + I_{ijl} + \mathrm{I}_{ikl} + I_{jkl} + I_{sijk} + I_{sijl} \\ &+ I_{sikl} + I_{sjkl} + I_{ijkl} + I_{sikjl} + z_{sijklt} \end{aligned}\right] \qquad (13.09)$$

where the subscripts s, i, j, k, l, and t refer to sex, sight, weight, rate, date of the particular tth individual, respectively; A is the grand mean of all individuals; B, C, D, E, and F are the measures of the main effects with respect to their own subscripts; the I's are the measures of interactions with respect to their own subscripts; and z_{sijklt} is experimental error.

The mathematical solution of the problem for securing the maximum likelihood estimates of each of the components in (13.09) is the same as that used in Chapter XI. In order to save space, we shall simply summarize all the results given in Table 90.

We wish to evaluate the 33 terms (listed below) in order to obtain all the sums of squares for the complete analysis of variance. To get the value of the term $\sum_s \sum_i \sum_j \sum_k \sum_l \sum_t X^2_{sijklt}$, we simply work out the sums of squares of all the figures in Table 89. There are two methods in evaluating each of the other terms. The first method includes three steps: (1) work out the squares for each sum of scores in the appropriate table;[2] (2) add the squares; (3) divide by the appropriate number which refers to the individual measures involved in each sum of scores. The second method also includes three steps: (1) work out the square for each mean score in the appropriate table;[2] (2) add these squares; (3) multiply

[2] "Appropriate table" refers to the table set up for securing the sum of scores and mean of scores required in each case. Since there were 37 tables required, they are not reproduced here. We shall illustrate the procedure in obtaining the sum of scores and the mean of scores for each subgroup. The sum of scores is obtained by adding the scores of (1) and (2) of Table 89. Thus: $4.5 + 14.0 = 18.5$. The mean score is obtained by dividing the sum of scores by 2: $\frac{18.5}{2} = 9.25$. Mathematically, the sum of scores is denoted by $\sum_t X_{sijklt}$, where $\sum_t$ means the summation of the two individuals; and the mean score is denoted by $\bar{X}_{sijkl\cdot}$, which is the mean score of the sth sex and the ith sight for the jth weight and the kth rate on the lth date.

TABLE 90

ANALYSIS OF VARIANCE OF D.L.-VALUES FOR THE FACTORIAL DESIGN 4 × 7 × 2 × 2 × 2

Source of variation	D.F.	Sum of squares
Error	224	$a - b$
Interaction		
4th order		
Sex × sight × weight × rate × date	18	$b - \sum_{p=1}^{5} c_p + \sum_{p=1}^{10} d_p - \sum_{p=1}^{10} e_p + \sum_{p=1}^{5} f_p - g$
3rd order		
Sex × sight × weight × rate	18	$c_1 - (d_1 + d_2 + d_4 + d_7) + (e_1 + e_2 + e_3 + e_4 + e_5 + e_6 + e_8)$ $- (f_1 + f_2 + f_3 + f_4) + g$
Sex × sight × weight × date	6	$c_2 - (d_1 + d_3 + d_5 + d_8) + (e_1 + e_2 + e_4 + e_5 + e_7 + e_9)$ $- (f_1 + f_2 + f_3 + f_5) + g$
Sex × sight × rate × date	3	$c_3 - (d_2 + d_3 + d_6 + d_9) + (e_1 + e_3 + e_4 + e_6 + e_7 + e_{10})$ $- (f_1 + f_2 + f_4 + f_5) + g$
Sex × weight × rate × date	18	$c_4 - (d_4 + d_5 + d_6 + d_{10}) + (e_2 + e_3 + e_4 + e_8 + e_9 + e_{10})$ $- (f_1 + f_3 + f_4 + f_5) + g$
Sight × weight × rate × date	18	$c_5 - (d_7 + d_8 + d_9 + d_{10}) + (e_5 + e_6 + e_7 + e_8 + e_9 + e_{10})$ $- (f_2 + f_3 + f_4 + f_5) + g$
2d order		
Sex × sight × weight	6	$d_1 - (e_1 + e_2 + e_5) + (f_1 + f_2 + f_3) - g$
Sex × sight × rate	3	$d_2 - (e_1 + e_3 + e_6) + (f_1 + f_2 + f_4) - g$
Sex × sight × date	1	$d_3 - (e_1 + e_4 + e_7) + (f_1 + f_2 + f_5) - g$
Sex × weight × rate	18	$d_4 - (e_2 + e_3 + e_8) + (f_1 + f_3 + f_4) - g$
Sex × weight × date	6	$d_5 - (e_2 + e_4 + e_9) + (f_1 + f_3 + f_5) - g$
Sex × rate × date	3	$d_6 - (e_3 + e_4 + e_{10}) + (f_1 + f_4 + f_5) - g$
Sight × weight × rate	18	$d_7 - (e_5 + e_6 + e_8) + (f_2 + f_3 + f_4) - g$
Sight × weight × date	6	$d_8 - (e_5 + e_7 + e_9) + (f_2 + f_3 + f_5) - g$
Sight × rate × date	3	$d_9 - (e_6 + e_7 + e_{10}) + (f_2 + f_4 + f_5) - g$
Weight × rate × date	18	$d_{10} - (e_8 + e_9 + e_{10}) + (f_3 + f_4 + f_5) - g$

TABLE 90 (*Continued*)

Source of variation	D.F.	Sum of squares
1st order		
Sex × sight	1	$e_1 - (f_1 + f_2) + g$
Sex × weight	6	$e_2 - (f_1 + f_3) + g$
Sex × rate	3	$e_3 - (f_1 + f_4) + g$
Sex × date	1	$e_4 - (f_1 + f_5) + g$
Sight × weight	6	$e_5 - (f_2 + f_3) + g$
Sight × rate	3	$e_6 - (f_2 + f_4) + g$
Sight × date	1	$e_7 - (f_2 + f_5) + g$
Weight × rate	18	$e_8 - (f_3 + f_4) + g$
Weight × date	6	$e_9 - (f_3 + f_5) + g$
Rate × date	3	$e_{10} - (f_4 + f_5) + g$
Main effects		
Sex	1	$f_1 - g$
Sight	1	$f_2 - g$
Weight	6	$f_3 - g$
Rate	3	$f_4 - g$
Date	1	$f_5 - g$
Total	447	$a - g$

by the appropriate number which refers to the individual measures involved in each mean score (this number will be the same as in the first method). We prefer to use the first method in calculation since it is more accurate from the viewpoint of significant figures. We use the second method, since it is simpler, in the presentation of the formulas.

By following the working procedure indicated in method 1, the values of all the 33 terms for our problem are obtained as follows:

$$a = \sum_s \sum_i \sum_j \sum_k \sum_l \sum_t X^2_{sijklt} = 441{,}140.30$$

$$b = \frac{\sum_s \sum_i \sum_j \sum_k \sum_l \left(\sum_t X_{sijklt}\right)^2}{2} = 2 \sum_s \sum_i \sum_j \sum_k \sum_l (\bar{X}^2_{sijkl.}) = 406{,}702.49$$

$$c_1 = \frac{\sum_s \sum_i \sum_j \sum_k \left(\sum_l \sum_t X_{sijklt}\right)^2}{4} = 4 \sum_s \sum_i \sum_j \sum_k (\bar{X}^2_{sijk..}) = 402{,}722.52$$

$$c_2 = \frac{\sum_s \sum_i \sum_j \sum_l \left(\sum_k \sum_t X_{sijklt}\right)^2}{8} = 8 \sum_s \sum_i \sum_j \sum_l (\bar{X}^2_{sij.l.}) = 342{,}295.66$$

$$c_3 = \frac{\sum_s \sum_i \sum_k \sum_l \left(\sum_j \sum_t X_{sijklt}\right)^2}{14} = 14 \sum_s \sum_i \sum_k \sum_l (\bar{X}^2_{si.kl.}) = 395{,}929.74$$

$$c_4 = \frac{\sum_s \sum_j \sum_k \sum_l \left(\sum_i \sum_t X_{sijklt}\right)^2}{4} = 4 \sum_s \sum_j \sum_k \sum_l (\bar{X}^2_{s.jkl.}) = 370{,}451.44$$

$$c_5 = \frac{\sum_i \sum_j \sum_k \sum_l \left(\sum_s \sum_t X_{sijklt}\right)^2}{4} = 4 \sum_i \sum_j \sum_k \sum_l (\bar{X}^2_{.ijkl.}) = 348{,}532.41$$

$$d_1 = \frac{\sum_s \sum_i \sum_j \left(\sum_k \sum_l \sum_t X_{sijklt}\right)^2}{16} = 16 \sum_s \sum_i \sum_j (\bar{X}^2_{sij...}) = 340{,}052.38$$

$$d_2 = \frac{\sum_s \sum_i \sum_k \left(\sum_j \sum_l \sum_t X_{sijklt}\right)^2}{28} = 28 \sum_s \sum_i \sum_k (\bar{X}^2_{si.k..}) = 395{,}333.63$$

$$d_3 = \frac{\sum_s \sum_i \sum_l \left(\sum_j \sum_k \sum_t X_{sijklt}\right)^2}{56} = 56 \sum_s \sum_i \sum_l (\bar{X}^2_{si..l.}) = 334{,}642.70$$

$$d_4 = \frac{\sum_s \sum_j \sum_k \left(\sum_s \sum_l \sum_t X_{sijklt}\right)^2}{8} = 8 \sum_s \sum_j \sum_k (\bar{X}^2_{s.jk..}) = 368{,}014.35$$

$$d_5 = \frac{\sum_s \sum_j \sum_l \left(\sum_i \sum_k \sum_t X_{sijklt}\right)^2}{16} = 16 \sum_s \sum_j \sum_l (\bar{X}^2_{s.j.l.}) = 311{,}866.98$$

$$d_6 = \frac{\sum_s \sum_k \sum_l \left(\sum_i \sum_j \sum_t X_{sijklt}\right)^2}{28} = 28 \sum_s \sum_k \sum_l (\bar{X}^2_{s..kl.}) = 365{,}342.11$$

$$d_7 = \frac{\sum_i \sum_j \sum_k \left(\sum_s \sum_l \sum_t X_{sijklt}\right)^2}{8} = 8 \sum_i \sum_j \sum_k (\bar{X}^2_{.ijk..}) = 346{,}776.80$$

$$d_8 = \frac{\sum_i \sum_j \sum_l \left(\sum_s \sum_k \sum_t X_{sijklt}\right)^2}{16} = 16 \sum_i \sum_j \sum_l (\bar{X}^2_{.ij.l.}) = 293{,}525.33$$

$$d_9 = \frac{\sum_i \sum_k \sum_l \left(\sum_s \sum_j \sum_t X_{sijklt}\right)^2}{28} = 28 \sum_i \sum_k \sum_l (\bar{X}^2_{.i.kl.}) = 341{,}930.40$$

$$d_{10} = \frac{\sum_j \sum_k \sum_l \left(\sum_s \sum_i \sum_t X_{sijklt}\right)^2}{8} = 8 \sum_j \sum_k \sum_l (\bar{X}^2_{..jkl.}) = 331{,}333.72$$

$$e_1 = \frac{\sum_s \sum_i \left(\sum_j \sum_k \sum_l \sum_t X_{sijklt}\right)^2}{112} = 112 \sum_s \sum_i (\bar{X}^2_{si....}) = 334{,}243.15$$

$$e_2 = \frac{\sum_s \sum_j \left(\sum_i \sum_k \sum_l \sum_t X_{sijklt}\right)^2}{32} = 32 \sum_s \sum_j (\bar{X}^2_{s.j...}) = 310{,}616.80$$

$$e_3 = \frac{\sum_s \sum_k \left(\sum_i \sum_j \sum_l \sum_t X_{sijklt}\right)^2}{56} = 56 \sum_s \sum_k (\bar{X}^2_{s..k..}) = 365{,}094.64$$

$$e_4 = \frac{\sum_s \sum_l \left(\sum_i \sum_j \sum_k \sum_t X_{sijklt}\right)^2}{112} = 112 \sum_s \sum_l (\bar{X}^2_{s...l.}) = 308{,}428.40$$

$$e_5 = \frac{\sum_i \sum_j \left(\sum_s \sum_k \sum_l \sum_t X_{sijklt}\right)^2}{32} = 32 \sum_i \sum_j (\bar{X}^2_{.ij...}) = 292{,}577.43$$

$$e_6 = \frac{\sum_i \sum_k \left(\sum_s \sum_j \sum_l \sum_t X_{sijklt}\right)^2}{56} = 56 \sum_i \sum_k (\bar{X}^2_{.i.k..}) = 341{,}734.41$$

$$e_7 = \frac{\sum_i \sum_l \left(\sum_s \sum_j \sum_k \sum_t X_{sijklt}\right)^2}{112} = 112 \sum_i \sum_l (\bar{X}^2_{.i..l.}) = 288{,}699.87$$

$$e_8 = \frac{\sum_j \sum_k \left(\sum_s \sum_i \sum_l \sum_t X_{sijklt}\right)^2}{16} = 16 \sum_j \sum_k (\bar{X}^2_{..jk..}) = 330{,}141.20$$

$$e_9 = \frac{\sum_j \sum_l \left(\sum_s \sum_i \sum_k \sum_t X_{sijklt}\right)^2}{32} = 32 \sum_j \sum_l (\bar{X}^2_{..j.l.}) = 279{,}282.48$$

$$e_{10} = \frac{\sum\limits_{k}\sum\limits_{l}\Big(\sum\limits_{s}\sum\limits_{i}\sum\limits_{j}\sum\limits_{t} X_{sijklt}\Big)^2}{56} = 56\sum_{k}\sum_{l}(\bar{X}^2_{\cdot\cdot\cdot kl\cdot}) = 328{,}018.72$$

$$f_1 = \frac{\sum\limits_{s}\Big(\sum\limits_{i}\sum\limits_{j}\sum\limits_{k}\sum\limits_{l}\sum\limits_{t} X_{sijklt}\Big)^2}{224} = 224\sum_{s}(\bar{X}^2_{s\cdot\cdot\cdot\cdot\cdot}) = 308{,}303.02$$

$$f_2 = \frac{\sum\limits_{i}\Big(\sum\limits_{s}\sum\limits_{j}\sum\limits_{k}\sum\limits_{l}\sum\limits_{t} X_{sijklt}\Big)^2}{224} = 224\sum_{i}(\bar{X}^2_{\cdot i\cdot\cdot\cdot\cdot}) = 288{,}579.46$$

$$f_3 = \frac{\sum\limits_{j}\Big(\sum\limits_{s}\sum\limits_{i}\sum\limits_{k}\sum\limits_{l}\sum\limits_{t} X_{sijklt}\Big)^2}{64} = 64\sum_{j}(\bar{X}^2_{\cdot\cdot j\cdot\cdot\cdot}) = 278{,}678.28$$

$$f_4 = \frac{\sum\limits_{k}\Big(\sum\limits_{s}\sum\limits_{i}\sum\limits_{j}\sum\limits_{l}\sum\limits_{t} X_{sijklt}\Big)^2}{112} = 112\sum_{k}(\bar{X}^2_{\cdot\cdot\cdot k\cdot\cdot}) = 327{,}841.31$$

$$f_5 = \frac{\sum\limits_{l}\Big(\sum\limits_{s}\sum\limits_{i}\sum\limits_{j}\sum\limits_{k}\sum\limits_{t} X_{sijklt}\Big)^2}{224} = 224\sum_{l}(\bar{X}^2_{\cdot\cdot\cdot\cdot l\cdot}) = 276{,}885.61$$

$$g = \frac{\Big(\sum\limits_{s}\sum\limits_{i}\sum\limits_{j}\sum\limits_{k}\sum\limits_{l}\sum\limits_{t} X_{sijklt}\Big)^2}{448} = 448\bar{X}^2_{\cdot\cdot\cdot\cdot\cdot\cdot} = 276{,}769.37$$

Substituting the above values in the appropriate formulas of Table 90, we obtain the specific sums of squares necessary for the complete analysis of variance.

We first test the significance of each of the interactions[3] of which there are 10 of the first order, 10 of the second, 5 of the third, and 1 of the fourth order. It is customary to call the interaction involving 2 factors an interaction of the first order; one involving 3 factors, 1 of the second order, and so on. The test of the significance of these interactions is given in Table 91. It is noted that the following interactions were significant:

sex × sight × rate	sight × rate
sex × sight	sight × weight (doubtful)
sex × rate	

The significant (including the doubtful) interactions were retained as specific components in the analysis-of-variance table. The statistically non-significant interactions were incorporated in experimental error.

The complete analysis of variance and the results of the corresponding tests of the respective hypotheses are given in Table 92.

[3] When two or more factors are involved such that increases or decreases in one (or more) influence increases or decreases in the other(s), or vice versa, *interaction* is said to exist.

TABLE 91
TESTS OF SIGNIFICANCE OF INTERACTIONS AS SOURCES OF VARIATION

Source of variation	D.F.	Sum of squares	Mean square	*F*	Test of hypothesis*
Error	224	34,438	154	.....	....
Sex × sight × weight × rate × datc.	18	270	15	.....	Accepted
Sex × sight × weight × rate	18	320	18	.....	Accepted
Sex × sight × weight × date	6	379	63	.....	Accepted
Sex × sight × rate × date	3	60	20	.....	Accepted
Sex × weight × rate × date	18	538	30	.....	Accepted
Sight × weight × rate × date	18	205	11	.....	Accepted
Sex × sight × weight	6	1,406	234	1.52	Accepted
Sex × sight × rate	3	2,216	739	4.80	Rejected
Sex × sight × date	1	270	270	1.75	Accepted
Sex × weight × rate	18	215	12	.....	Accepted
Sex × weight × date	6	637	106	.....	Accepted
Sex × rate × date	3	61	20	.....	Accepted
Sight × weight × rate	18	654	36	.....	Accepted
Sight × weight × date	6	340	57	.....	Accepted
Sight × rate × date	3	14	5	.....	Accepted
Weight × rate × date	18	527	29	.....	Accepted
Sex × sight	1	14,130	14,130	91.75	Rejected
Sex × weight	6	405	68	.....	Accepted
Sex × rate	3	5,720	1,907	12.38	Rejected
Sex × date	1	9	9	.....	Accepted
Sight × weight	6	2,089	348	2.26	Remains in doubt
Sight × rate	3	2,083	694	4.51	Rejected
Sight × date	1	4	4	.....	Accepted
Weight × rate	18	391	22	.....	Accepted
Weight × date	6	488	81	.....	Accepted
Rate × date	3	61	20	.....	Accepted

* The hypothesis tested is a null hypothesis concerning the variation in the samc row. For example, the hypothesis regarding sex × sight × weight × rate × date is that there is no significant interaction between sex, sight, weight, rate, and datc.

The tests of significance resulted in the following conclusions:

significant main effects: sex, sight, weight, and rate
significant second-order interactions: sex × sight × rate
significant first-order interactions: sex × sight
sex × rate
sight × weight
sight × rate

It is worth noting that there was no significant difference between dates and that no interaction including date as a component was significant. This result demonstrates that the observations were consistent among themselves.

TABLE 92
COMPLETE ANALYSIS OF VARIANCE OF D.L.-VALUES

Source of variation	D.F.	Sum of squares	Mean square	*F*	Test of hypothesis*
Residual..................	419	41,692	100	......	....
Sex × sight × rate........	3	2,216	739	7.39	Rejected
Sex × sight...............	1	14,130	14,130	141.30	Rejected
Sex × rate................	3	5,720	1,907	19.07	Rejected
Sight × weight............	6	2,089	348	3.48	Rejected
Sight × rate..............	3	2,083	694	6.94	Rejected
Sex.......................	1	31,534	31,534	315.34	Rejected
Sight.....................	1	11,810	11,810	118.10	Rejected
Weight....................	6	1,909	318	3.18	Rejected
Rate......................	3	51,072	17,024	170.24	Rejected
Date......................	1	116	116	1.16	Accepted
Total.................	447	164,371			

* The hypothesis tested is a null hypothesis regarding the variation in the same row. For example, the hypothesis concerning date is that there is no significant difference between the date means.

From the standpoint of the efficiency of the factorial design in this experiment, it can be said that we have tested 26 hypotheses regarding interactions and 5 hypotheses concerning main effects. If we had used the single-factor plan of experiment, we should have required 56 experiments for testing the main effects of rate; 32, for weight; 112, for sex; 112, for sight; and 112, for date. We also would have had to repeat the *t*-test for $C_2^{4\times7\times2\times2\times2} = C_2^{224}$ times. Furthermore, no information would be possible concerning the interaction effects.

The Problem of Prediction. The regression equations of D.L.-values on each of the factors and interacting factors, which were found to be significant, can be determined. With these equations it is possible to compute D.L.-values for any particular value of the independent variable within the range of factor levels used in the experiment.

We shall illustrate the use of orthogonal polynomials for determining the regression equation for predicting D.L.-values from weights.[4]

We proceed to work out linear, quadratic, and cubic regression equations. Only the linear coefficient was found significant here, but the methods of calculating the latter two are also given. We shall show the

[4] For the regression equations of the other significant factors in this study, see the original article, Ref. 14. Other useful references are 2 and 10, particularly 10, for the discussion of the meaning of the linear, quadratic, and cubic terms.

method of separating effects associated with more than one degree of freedom into component parts that are mutually orthogonal. Because of the latter property, the components may be estimated from the data. If in our experiment there is only 1 degree of freedom representing the tested variation, for example, sex and date, there can be only a linear relation between the two levels of variation and the D.L.-values. If there are more than 2 degrees of freedom or more than 3 levels of variation, then these can be separated into component parts—linear, quadratic, cubic, and so on—that are mutually independent. Even when there are more than 3 degrees of freedom or more than 4 levels of variation, we usually do not calculate terms higher than the cubic.

We first record the means of the D.L.-values found for each weight and transform them as follows:

W (weight)	Y (D.L.-value)	x	y
100	$28.792\bar{2}$	*−3*	3.9368
150	$26.414\bar{1}$	*−2*	1.5587
200	$25.634\bar{4}$	*−1*	0.7790
250	$23.745\dot{3}$	0	*−1.1100*
300	$23.829\bar{7}$	1	*−1.0257*
350	$23.256\bar{3}$	2	*−1.5991*
400	$22.315\dot{6}$	3	*−2.5397*
$\bar{W} = 250$	$\bar{Y} = 24.855\bar{4}$	$\Sigma x^2 = 28$	

where $x = \dfrac{W - 250}{50}$, $y = Y - 24.8554.$

We then refer to the tables of Fisher and Yates on orthogonal polynomials (Ref. 8) for $N = 7$, which reads:

ξ_1'	ξ_2'	ξ_3'
−3	5	*−1*
−2	0	1
−1	*−3*	1
0	*−4*	0
1	*−3*	*−1*
2	0	*−1*
3	5	1
$\Sigma\xi_1'^2 = 28$	$\Sigma\xi_2'^2 = 84$	$\Sigma\xi_3'^2 = 6$
$\lambda_1 = 1$	$\lambda_2 = 1$	$\lambda_3 = \frac{1}{6}$

Finally, we obtain all the regression equations as follows:

Linear: $$\hat{Y} = c_0 + c_1x \tag{13.10}$$

where $c_0 = \bar{Y} = 24.8554$

$$c_1 = \frac{\Sigma \xi_1' y}{\Sigma \xi_1'^2} \lambda_1$$

Quadratic:
$$\hat{Y} = c_0' + c_1 x + c_2 x^2 \tag{13.11}$$
where
$$c_0' = c_0 - \frac{(\Sigma x^2) c_2}{n}$$

Cubic:
$$\hat{Y} = c_0' + c_1 x + c_2 x^2 + c_3 x^3 \tag{13.12}$$
where
$$c_0 = \bar{Y} \tag{13.13}$$
$$c_0' = c_0 - \frac{(\Sigma x^2) c_2}{n} \tag{13.14}$$
$$c_1 = \frac{\Sigma \xi_1' y}{\Sigma \xi_1'^2} \lambda_1 \tag{13.15}$$
$$c_2 = \frac{\Sigma \xi_2' y}{\Sigma \xi_2'^2} \lambda_2 \tag{13.16}$$
$$c_3 = \frac{\Sigma \xi_3' y}{\Sigma \xi_3'^2} \lambda_3 \tag{13.17}$$

The calculation of the regression coefficients for weights is carried out as follows:

x	y	ξ_1'	$\xi_1' y$	ξ_2'	$\xi_2' y$	ξ_3'	$\xi_3' y$
−3	3.9368	−3	−11.8104	5	19.6840	−1	−3.9368
−2	1.5587	−2	− 3.1174	0	0.0000	1	1.5587
−1	0.7790	−1	− 0.7790	−3	− 2.3370	1	0.7790
0	−1.1100	0	0.0000	−4	4.4400	0	0.0000
1	−1.0257	1	− 1.0257	−3	3.0771	−1	1.0257
2	−1.5991	2	− 3.1982	0	0.0000	−1	1.5991
3	−2.5397	3	− 7.6191	5	−12.6985	1	−2.5397
		$\Sigma \xi_1'^2 = 28$	$\Sigma \xi_1' y = -27.5498$	$\Sigma \xi_2'^2 = 84$	$\Sigma \xi_2' y = 12.1656$	$\Sigma \xi_3'^2 = 6$	$\Sigma \xi_3' y = 1.5140$
		$\lambda_1 = 1,$		$\lambda_2 = 1,$		$\lambda_3 = \frac{1}{6}$	

By using Equations (13.13), (13.14), (13.15), (13.16), and (13.17), we obtain

$$c_0 = 24.8554 \qquad c_2 = .144829$$
$$c_1 = -.983921 \qquad c_3 = -.042056$$
$$c_0' = 24.2761$$

Hence, the regression equations can be obtained by substituting these values into Equations (13.10), (13.11), and (13.12).

Linear: $\hat{Y} = 24.8554 - .983921x$ (13.18)

Quadratic: $\hat{Y} = 24.2761 - .983921x + .144829x^2$ (13.19)

Cubic: $\hat{Y} = 24.2761 - .983921x + .144829x^2 - .042056x^3$ (13.20)

where $x = \dfrac{W - 250}{50}$.

The test of significance of the components of variation due to weight is given in Table 93.

TABLE 93
COMPONENTS OF VARIATION DUE TO WEIGHT

Source of variation	D.F.	Sum of squares	Mean square	F	Test of hypothesis
Linear	1	1735	1735	17.35	Rejected
Quadratic	1	113	113	1.13	Accepted
Cubic	1	24	24	.....	Accepted
Remainder	3	37	12	.....	Accepted
Weights	6	1909			

It is noted from Table 93 that only the linear component is significant. Hence, only the linear equation is to be used in prediction. The graph of the linear regression equation for the observed D.L.-values is sketched in Fig. 8.

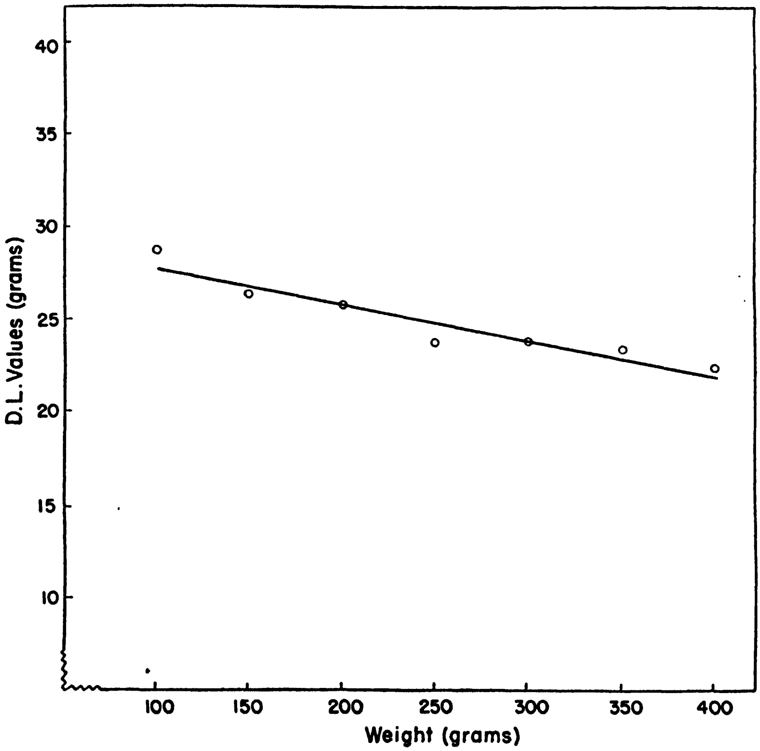


Figure 8. Linear regression line of the equation for predicting D. L. values from weight values.

Factorial Design and Covariance in a Study of Educational Development. We wish to illustrate further application of the principles of factorial design by presenting the results of an investigation of individual educational development.[5] An application is also made in this study of the method of covariance which served to increase the precision of the experiment. The specific design developed for this study was a 2 × 3 × 3 × 3 factorial type. The factors chosen for study were the 2 sexes, 3 scholastic standings, 3 individual orders, and three school grades.

In addition to the introduction of the covariance method for controlling variables not controlled or controllable directly by the experimental design, this experiment differs from the one in psychology just reported in that the type of factorial design is of the kind in which absolute replication is dispensed with and hidden replication is involved (Ref. 7). This type is desirable when large numbers of combinations are tested simultaneously without repeated use of each combination. All the independent comparisons contained in the experiment are allotted to the factors tested and to their interactions. Since there is no independent comparison ascribable to pure error, the highest order interactions are employed as the basis for measuring the precision of the main comparisons. The situation in this study has a very wide occurrence in research work.

The criterion score used as a measure of the stage of educational development was based on a composite score comprised of the scores on nine separate tests (Ref. 13). The standard scores used—ranging from 0 to 30 with a mean of 15—were determined from the combined grades, that is, the tenth, eleventh, and twelfth grades. There were 18 students from each of the 3 grades, all chosen at random from the total number enrolled in these grades. The mental-age scores were obtained from the administration of a group test of mental ability and were calculated for all students as of the same date. All students in the tenth grade were of chronological age fifteen; in grade 11, sixteen; in grade 12, seventeen. Students were classified into one of three scholastic groups—good, average, poor—based on their honor-point ratios. Individual order of educational development was based on the size of the scores of the individuals on the second of the two administrations of the battery of tests. The interval between the two administrations was 12 months.

Let us denote the final score, the initial score, and the mental-age score by Y, X_1, and X_2, respectively. Again, the two sexes are denoted by I for the male and II for the female; three grades are denoted by A for grade 10; B, for grade 11; and C, for grade 12. The three scholastic standings are denoted by 1 for the good, 2 for the average, and 3 for the poor; and the three individual orders by α for the first, β for the second, and γ for the third. The primary data grouped into the several sub-

[5] For the complete analysis of the experimental results in this investigation, their interpretation, and the mathematical formulation and solution of the problem, read Ref. 13.

classes in accordance with the notations specified are presented in Table 94.

TABLE 94

SCORES FOR ALL SEX × GRADE × SCHOLASTIC × INDIVIDUAL COMBINATIONS

		Grade	A			B			C		
Sex	Scholastic standard	Individual / Measure	Y	X_1	X_2	Y	X_1	X_2	Y	X_1	X_2
I	1	α	30	28	45	26	22	62	29	25	60
		β	25	22	58	26	21	57	29	24	88
		γ	22	19	46	24	21	65	22	19	64
	2	α	26	22	56	24	25	54	23	21	64
		β	17	14	19	23	18	55	20	17	47
		γ	14	14	29	15	13	24	19	17	75
	3	α	18	18	34	18	17	40	17	16	29
		β	17	14	17	16	13	24	15	15	38
		γ	12	9	19	13	12	23	14	12	28
II	1	α	21	16	44	26	22	60	33	29	94
		β	21	21	44	25	22	57	29	29	89
		γ	19	17	6	23	19	52	25	22	78
	2	α	20	18	38	22	19	54	23	21	50
		β	18	16	27	21	19	54	18	19	57
		γ	14	14	18	17	16	52	17	17	43
	3	α	14	9	18	19	17	40	15	13	36
		β	12	7	18	15	12	28	15	14	35
		γ	9	7	5	13	12	48	10	9	14

In our problem, we define:

Y_{sijt} = the final standard score of the tth individual of the jth scholastic standing in the ith grade and the sth sex.

$X_{1_{sijt}}$ = the initial standard score of the tth individual of the jth scholastic standing in the ith grade and the sth sex.

$X_{2_{sijt}}$ = the mental-age score of the tth individual of the jth standing in the ith grade and the sth sex.

In the above definitions, $s = 1, 2$; $i = 1, 2, 3$; $j = 1, 2, 3$; $t = 1, 2, 3$. We then proceed to obtain all the sum of squares and products required for the analysis as shown in Table 95. These are listed below, together with the notation for each quantity. We shall illustrate how these values are obtained by two examples.

EXAMPLE 1. In order to evaluate the term $\sum_s \sum_i \sum_j \sum_t Y^2_{sijt}$, we simply refer to Table 94 and work out all the squares of the Y-measures. Then

TABLE 95

SUM OF SQUARES AND OF PRODUCTS FOR EACH SOURCE OF VARIATION

Source of variation	D.F.	Σy^2	Σx_1^2	Σx_2^2	$\Sigma x_1 y$	$\Sigma x_2 y$	$\Sigma x_1 x_2$
Sex × grade × schol. × individual	8	$a_1 - \sum_{k=1}^{4} b_{1k} + \sum_{k=1}^{6} c_{1k} - \sum_{k=1}^{4} d_{1k} + e_1$	$a_2 - \sum_{k=1}^{4} b_{2k} + \sum_{k=1}^{6} c_{2k} - \sum_{k=1}^{4} d_{2k} + e_2$	$a_3 - \sum_{k=1}^{4} b_{3k} + \sum_{k=1}^{6} c_{3k} - \sum_{k=1}^{4} d_{3k} + e_3$	$a_4 - \sum_{k=1}^{4} b_{4k} + \sum_{k=1}^{6} c_{4k} - \sum_{k=1}^{4} d_{4k} + e_4$	$a_5 - \sum_{k=1}^{4} b_{5k} + \sum_{k=1}^{6} c_{5k} - \sum_{k=1}^{4} d_{5k} + e_5$	$a_6 - \sum_{k=1}^{4} b_{6k} + \sum_{k=1}^{6} c_{6k} - \sum_{k=1}^{4} d_{6k} + e_6$
Sex × grade × schol.	4	$b_{11} - (c_{11} + c_{12} + c_{14}) + (d_{11} + d_{12} + d_{13}) - e_1$	$b_{21} - (c_{21} + c_{22} + c_{24}) + (d_{21} + d_{22} + d_{23}) - e_2$	$b_{31} - (c_{31} - c_{32} + c_{34}) + (d_{31} + d_{32} + d_{33}) - e_3$	$b_{41} - (c_{41} + c_{42} + c_{44}) + (d_{41} + d_{42} + d_{43}) - e_4$	$b_{51} - (c_{51} + c_{52} + c_{54}) + (d_{51} + d_{52} + d_{53}) - e_5$	$b_{61} - (c_{61} + c_{62} + c_{64}) + (d_{61} + d_{62} + d_{63}) - e_6$
Sex × grade × individual	4	$b_{12} - (c_{11} + c_{13} + c_{15}) + (d_{11} + d_{12} + d_{14}) - e_1$	$b_{22} - (c_{21} + c_{23} + c_{25}) + (d_{21} + d_{22} + d_{24}) - e_2$	$b_{32} - (c_{31} + c_{33} + c_{35}) + (d_{31} + d_{32} + d_{34}) - e_3$	$b_{42} - (c_{41} + c_{43} + c_{45}) + (d_{41} + d_{42} + d_{44}) - e_4$	$b_{52} - (c_{51} + c_{53} + c_{55}) + (d_{51} + d_{52} + d_{54}) - e_5$	$b_{62} - (c_{61} + c_{63} + c_{65}) + (d_{61} + d_{62} + d_{64}) - e_6$
Sex × schol. × individual	4	$b_{13} - (c_{12} + c_{13} + c_{16}) + (d_{11} + d_{13} + d_{14}) - e_1$	$b_{23} - (c_{22} + c_{23} + c_{26}) + (d_{21} + d_{23} + d_{24}) - e_2$	$b_{33} - (c_{32} + c_{33} + c_{36}) + (d_{31} + d_{33} + d_{34}) - e_3$	$b_{43} - (c_{42} + c_{43} + c_{46}) + (d_{41} + d_{43} + d_{44}) - e_4$	$b_{53} - (c_{52} + c_{53} + c_{56}) + (d_{51} + d_{53} + d_{54}) - e_5$	$b_{63} - (c_{62} + c_{63} + c_{66}) + (d_{61} + d_{63} + d_{64}) - e_6$
Grade × schol. × individual	8	$b_{14} - (c_{14} + c_{15} + c_{16}) + (d_{12} + d_{13} + d_{14}) - e_1$	$b_{24} - (c_{24} + c_{25} + c_{26}) + (d_{22} + d_{23} + d_{24}) - e_2$	$b_{34} - (c_{34} + c_{35} + c_{36}) + (d_{32} + d_{33} + d_{34}) - e_3$	$b_{44} - (c_{44} + c_{45} + c_{46}) + (d_{42} + d_{43} + d_{44}) - e_4$	$b_{54} - (c_{54} + c_{55} + c_{56}) + (d_{52} + d_{53} + d_{54}) - e_5$	$b_{64} - (c_{64} + c_{65} + c_{66}) - (d_{62} + d_{63} + d_{64}) - e_6$
Sex × grade	2	$c_{11} - (d_{11} + d_{12}) + e_1$	$c_{21} - (d_{21} + d_{22}) + e_2$	$c_{31} - (d_{31} + d_{32}) + e_3$	$c_{41} - (d_{41} + d_{12}) + e_4$	$c_{51} - (d_{51} + d_{52}) + e_5$	$c_{61} - (d_{61} + d_{62}) + e_6$
Sex × schol.	2	$c_{12} - (d_{11} + d_{13}) + e_1$	$c_{22} - (d_{21} + d_{23}) + e_2$	$c_{32} - (d_{31} + d_{33}) + e_3$	$c_{42} - (d_{41} + d_{43}) + e_4$	$c_{52} - (d_{51} + d_{54}) + e_5$	$c_{62} - (d_{61} + d_{63}) + e_6$
Sex × individual	2	$c_{13} - (d_{11} + d_{14}) + e_1$	$c_{23} - (d_{21} + d_{24}) + e_2$	$c_{33} - (d_{31} + d_{34}) + e_3$	$c_{43} - (d_{41} + d_{44}) + e_4$	$c_{53} - (d_{51} + d_{54}) + e_5$	$c_{63} - (d_{61} + d_{64}) + e_6$
Grade × schol.	4	$c_{14} - (d_{12} + d_{13}) + e_1$	$c_{24} - (d_{22} + d_{23}) + e_2$	$c_{34} - (d_{32} + d_{33}) + e_3$	$c_{44} - (d_{42} + d_{43}) + e_4$	$c_{54} - (d_{52} + d_{53}) + e_5$	$c_{64} - (d_{62} + d_{63}) + e_6$
Grade × individual	4	$c_{15} - (d_{12} + d_{14}) + e_1$	$c_{25} - (d_{22} + d_{24}) + e_2$	$c_{35} - (d_{32} + d_{34}) + e_3$	$c_{45} - (d_{42} + d_{44}) + e_4$	$c_{55} - (d_{52} + d_{54}) + e_5$	$c_{65} - (d_{62} + d_{64}) + e_6$
Schol. × individual	4	$c_{16} - (d_{13} + d_{14}) + e_1$	$c_{26} - (d_{23} + d_{24}) + e_2$	$c_{36} - (d_{33} + d_{34}) + e_3$	$c_{46} - (d_{43} + d_{44}) + e_4$	$c_{56} - (d_{53} + d_{54}) + e_5$	$c_{66} - (d_{63} + d_{64}) + e_6$
Sex	1	$d_{11} - e_1$	$d_{21} - e_2$	$d_{31} - e_3$	$d_{41} - e_4$	$d_{51} - e_5$	$d_{61} - e_6$
Grade	2	$d_{12} - e_1$	$d_{22} - e_2$	$d_{32} - e_3$	$d_{42} - e_4$	$d_{52} - e_5$	$d_{62} - e_6$
Scholastic	2	$d_{13} - e_1$	$d_{23} - e_2$	$d_{33} - e_3$	$d_{43} - e_4$	$d_{53} - e_5$	$d_{63} - e_6$
Individual	2	$d_{14} - e_1$	$d_{24} - e_2$	$d_{34} - e_3$	$d_{44} - e_4$	$d_{54} - e_5$	$d_{64} - e_6$
Total	53	$a_1 - e_1$	$a_2 - e_2$	$a_3 - e_3$	$a_4 - e_4$	$a_5 - e_5$	$a_6 - e_6$

we sum these squares and obtain the required value, for example, $a_1 = 22{,}730$.

TABLE 96
SUM OF SCORES FOR EACH SEX × GRADE × SCHOLASTIC COMBINATION

	Grade	A			B			C		
Sex	Scholastic standard / Measure	ΣY	ΣX_1	ΣX_2	ΣY	ΣX_1	ΣX_2	ΣY	ΣX_1	ΣX_2
I	1	77	69	149	76	64	184	80	68	212
	2	57	50	104	62	56	133	62	55	186
	3	47	41	70	47	42	87	46	43	95
II	1	61	54	94	74	63	169	87	80	261
	2	52	48	83	60	54	160	58	57	150
	3	35	23	41	47	41	116	40	36	85

EXAMPLE 2. In order to evaluate the term

$$\sum_s \sum_i \sum_j \left[\frac{\left(\sum_t X_{1_{sijt}}\right)\left(\sum_t Y_{sijt}\right)}{3} \right],$$

we refer to Table 96, then compute all the products of ΣY and ΣX_1 in the same row. We then add these products and divide by 3 to obtain the quantity: $b_{41} = 19{,}786$.

The sum of scores for each sex × grade × scholastic combination as given in Table 96 was obtained by adding the scores for α, β, and γ as given in Table 94. Thus: $30 + 25 + 22 = 77$.

By following similar procedures as illustrated for Examples 1 and 2, we obtain all the values for the 96 terms extending from a_1 through e_6.

Here we shall present the results based on one analysis only:[6] the complete analysis of variance and covariance partialing out the effects of both initial score and mental age.

Values required for obtaining the sums of squares and products specified in Table 95:

$$a_1 = \sum_s \sum_i \sum_j \sum_t Y^2_{sijt} = 22{,}730$$

$$a_2 = \sum_s \sum_i \sum_j \sum_t X^2_{1_{sijt}} = 17{,}926$$

[6] For a complete analysis see Ref. 13. The examination of the assumptions underlying the analysis of variance and covariance led to their acceptance insofar as they could be tested. See pages 218–219 and Ref. 1 in Chapter X, and pages 251–260 in Chapter XI.

$$a_3 = \sum_s \sum_i \sum_j \sum_t X^2_{2_{sijt}} = 127{,}369$$

$$a_4 = \sum_s \sum_i \sum_j \sum_t (X_{1_{sijt}} Y_{sijt}) = 20{,}116$$

$$a_5 = \sum_s \sum_i \sum_j \sum_t (X_{2_{sijt}} Y_{sijt}) = 52{,}005$$

$$a_6 = \sum_s \sum_i \sum_j \sum_t (X_{1_{sijt}} X_{2_{sijt}}) = 46{,}227$$

$$b_{11} = \sum_s \sum_i \sum_j \left[\frac{\left(\sum_t Y_{sijt} \right)^2}{3} \right] = 22{,}348$$

$$b_{12} = \sum_s \sum_i \sum_t \left[\frac{\left(\sum_j Y_{sijt} \right)^2}{3} \right] = 21{,}565$$

$$b_{13} = \sum_s \sum_j \sum_t \left[\frac{\left(\sum_i Y_{sijt} \right)^2}{3} \right] = 22{,}500$$

$$b_{14} = \sum_i \sum_j \sum_t \left[\frac{\left(\sum_s Y_{sijt} \right)^2}{2} \right] = 22{,}599$$

$$b_{21} = \sum_s \sum_i \sum_j \left[\frac{\left(\sum_t X_{1_{sijt}} \right)^2}{3} \right] = 17{,}565$$

$$b_{22} = \sum_s \sum_i \sum_t \left[\frac{\left(\sum_j X_{1_{sijt}} \right)^2}{3} \right] = 16{,}939$$

$$b_{23} = \sum_s \sum_j \sum_t \left[\frac{\left(\sum_i X_{1_{sijt}} \right)^2}{3} \right] = 17{,}643$$

$$b_{24} = \sum_i \sum_j \sum_t \left[\frac{\left(\sum_s X_{1_{sijt}} \right)^2}{2} \right] = 17{,}712$$

$$b_{31} = \sum_s \sum_i \sum_j \left[\frac{\left(\sum_t X_{2_{sijt}} \right)^2}{3} \right] = 122{,}832$$

$$b_{32} = \sum_s \sum_i \sum_t \left[\frac{\left(\sum_j X_{2_{sijt}} \right)^2}{3} \right] = 113{,}596$$

$$b_{33} = \sum_s \sum_j \sum_t \left[\frac{\left(\sum_i X_{2_{sijt}} \right)^2}{3} \right] = 116{,}146$$

$$b_{34} = \sum_i \sum_j \sum_t \left[\frac{\left(\sum_s X_{2sijt}\right)^2}{2} \right] = 123{,}997$$

$$b_{41} = \sum_s \sum_i \sum_j \left[\frac{\left(\sum_t X_{1sijt}\right)\left(\sum_t Y_{sijt}\right)}{3} \right] = 19{,}786$$

$$b_{42} = \sum_s \sum_i \sum_t \left[\frac{\left(\sum_j X_{1sijt}\right)\left(\sum_j Y_{sijt}\right)}{3} \right] = 19{,}084$$

$$b_{43} = \sum_s \sum_j \sum_t \left[\frac{\left(\sum_i X_{1sijt}\right)\left(\sum_i Y_{sijt}\right)}{3} \right] = 19{,}891$$

$$b_{44} = \sum_i \sum_j \sum_t \left[\frac{\left(\sum_s X_{1sijt}\right)\left(\sum_s Y_{sijt}\right)}{2} \right] = 19{,}977$$

$$b_{51} = \sum_s \sum_i \sum_j \left[\frac{\left(\sum_t X_{2sijt}\right)\left(\sum_t Y_{sijt}\right)}{3} \right] = 51{,}241$$

$$b_{52} = \sum_s \sum_i \sum_t \left[\frac{\left(\sum_j X_{2sijt}\right)\left(\sum_j Y_{sijt}\right)}{3} \right] = 48{,}422$$

$$b_{53} = \sum_s \sum_j \sum_t \left[\frac{\left(\sum_i X_{2sijt}\right)\left(\sum_i X_{sijt}\right)}{3} \right] = 50{,}734$$

$$b_{54} = \sum_i \sum_j \sum_t \left[\frac{\left(\sum_s X_{2sijt}\right)\left(\sum_s Y_{sijt}\right)}{2} \right] = 51{,}639$$

$$b_{61} = \sum_s \sum_i \sum_j \left[\frac{\left(\sum_t X_{1sijt}\right)\left(\sum_t X_{2sijt}\right)}{3} \right] = 45{,}499$$

$$b_{62} = \sum_s \sum_i \sum_t \left[\frac{\left(\sum_j X_{1sijt}\right)\left(\sum_j X_{2sijt}\right)}{3} \right] = 43{,}010$$

$$b_{63} = \sum_s \sum_j \sum_t \left[\frac{\left(\sum_i X_{1sijt}\right)\left(\sum_i X_{2sijt}\right)}{3} \right] = 44{,}924$$

$$b_{64} = \sum_i \sum_j \sum_t \left[\frac{\left(\sum_s X_{1sijt}\right)\left(\sum_s X_{2sijt}\right)}{2} \right] = 45{,}882$$

$$c_{11} = \sum_s \sum_i \left[\frac{\left(\sum_j \sum_t Y_{sijt}\right)^2}{9} \right] = 21{,}247$$

$$c_{12} = \sum_{s}\sum_{j}\left[\frac{\left(\sum_{i}\sum_{t} Y_{sijt}\right)^2}{9}\right] = 22{,}191$$

$$c_{13} = \sum_{s}\sum_{t}\left[\frac{\left(\sum_{i}\sum_{j} Y_{sijt}\right)^2}{9}\right] = 21{,}447$$

$$c_{14} = \sum_{i}\sum_{j}\left[\frac{\left(\sum_{s}\sum_{t} Y_{sijt}\right)^2}{6}\right] = 22{,}259$$

$$c_{15} = \sum_{i}\sum_{t}\left[\frac{\left(\sum_{s}\sum_{j} Y_{sijt}\right)^2}{6}\right] = 21{,}482$$

$$c_{16} = \sum_{j}\sum_{t}\left[\frac{\left(\sum_{s}\sum_{i} Y_{sijt}\right)^2}{6}\right] = 22{,}461$$

$$c_{21} = \sum_{s}\sum_{i}\left[\frac{\left(\sum_{j}\sum_{t} X_{1sijt}\right)^2}{9}\right] = 16{,}658$$

$$c_{22} = \sum_{s}\sum_{j}\left[\frac{\left(\sum_{i}\sum_{t} X_{1sijt}\right)^2}{9}\right] = 17{,}365$$

$$c_{23} = \sum_{s}\sum_{t}\left[\frac{\left(\sum_{i}\sum_{j} X_{1sijt}\right)^2}{9}\right] = 16{,}774$$

$$c_{24} = \sum_{i}\sum_{j}\left[\frac{\left(\sum_{s}\sum_{t} X_{1sijt}\right)^2}{6}\right] = 17{,}439$$

$$c_{25} = \sum_{i}\sum_{t}\left[\frac{\left(\sum_{s}\sum_{j} X_{1sijt}\right)^2}{6}\right] = 16{,}813$$

$$c_{26} = \sum_{j}\sum_{t}\left[\frac{\left(\sum_{s}\sum_{i} X_{1sijt}\right)^2}{6}\right] = 17{,}564$$

$$c_{31} = \sum_{s}\sum_{i}\left[\frac{\left(\sum_{j}\sum_{t} X_{2sijt}\right)^2}{9}\right] = 111{,}351$$

$$c_{32} = \sum_{s}\sum_{j}\left[\frac{\left(\sum_{i}\sum_{t} X_{2sijt}\right)^2}{9}\right] = 114{,}116$$

$$c_{33} = \sum_{s}\sum_{t}\left[\frac{\left(\sum_{i}\sum_{j} X_{2sijt}\right)^2}{9}\right] = 106{,}019$$

$$c_{34} = \sum_i \sum_j \left[\frac{\left(\sum_s \sum_t X_{2sijt} \right)^2}{6} \right] = 121{,}172$$

$$c_{35} = \sum_i \sum_t \left[\frac{\left(\sum_s \sum_j X_{2sijt} \right)^2}{6} \right] = 112{,}112$$

$$c_{36} = \sum_j \sum_t \left[\frac{\left(\sum_s \sum_i X_{2sijt} \right)^2}{6} \right] = 115{,}432$$

$$c_{41} = \sum_s \sum_i \left[\frac{\left(\sum_j \sum_t X_{1sijt} \right) \left(\sum_j \sum_t Y_{sijt} \right)}{9} \right] = 18{,}805$$

$$c_{42} = \sum_s \sum_j \left[\frac{\left(\sum_i \sum_t X_{1sijt} \right) \left(\sum_i \sum_t Y_{sijt} \right)}{9} \right] = 19{,}620$$

$$c_{43} = \sum_s \sum_t \left[\frac{\left(\sum_i \sum_j X_{1sijt} \right) \left(\sum_i \sum_j Y_{sijt} \right)}{9} \right] = 18{,}954$$

$$c_{44} = \sum_i \sum_j \left[\frac{\left(\sum_s \sum_t X_{1sijt} \right) \left(\sum_s \sum_t Y_{sijt} \right)}{6} \right] = 19{,}688$$

$$c_{45} = \sum_i \sum_t \left[\frac{\left(\sum_s \sum_j X_{1sijt} \right) \left(\sum_s \sum_j Y_{sijt} \right)}{6} \right] = 18{,}994$$

$$c_{46} = \sum_j \sum_t \left[\frac{\left(\sum_s \sum_i X_{1sijt} \right) \left(\sum_s \sum_i Y_{sijt} \right)}{6} \right] = 19{,}853$$

$$c_{51} = \sum_s \sum_i \left[\frac{\left(\sum_j \sum_t X_{2sijt} \right) \left(\sum_j \sum_t Y_{sijt} \right)}{9} \right] = 47{,}828$$

$$c_{52} = \sum_s \sum_j \left[\frac{\left(\sum_i \sum_t X_{2sijt} \right) \left(\sum_i \sum_t Y_{sijt} \right)}{9} \right] = 50{,}166$$

$$c_{53} = \sum_s \sum_t \left[\frac{\left(\sum_i \sum_j X_{2sijt} \right) \left(\sum_i \sum_j Y_{sijt} \right)}{9} \right] = 47{,}627$$

$$c_{54} = \sum_i \sum_j \left[\frac{\left(\sum_s \sum_t X_{2sijt} \right) \left(\sum_s \sum_t Y_{sijt} \right)}{6} \right] = 50{,}931$$

$$c_{55} = \sum_i \sum_t \left[\frac{\left(\sum_s \sum_j X_{2sijt} \right) \left(\sum_s \sum_j Y_{sijt} \right)}{6} \right] = 48{,}212$$

$$c_{56} = \sum_j \sum_t \left[\frac{\left(\sum_s \sum_i X_{2sijt}\right)\left(\sum_s \sum_i Y_{sijt}\right)}{6} \right] = 50{,}698$$

$$c_{61} = \sum_s \sum_i \left[\frac{\left(\sum_j \sum_t X_{1sijt}\right)\left(\sum_j \sum_t X_{2sijt}\right)}{9} \right] = 42{,}482$$

$$c_{62} = \sum_s \sum_j \left[\frac{\left(\sum_i \sum_t X_{1sijt}\right)\left(\sum_i \sum_t X_{2sijt}\right)}{9} \right] = 44{,}368$$

$$c_{63} = \sum_s \sum_t \left[\frac{\left(\sum_i \sum_j X_{1sijt}\right)\left(\sum_i \sum_j X_{2sijt}\right)}{9} \right] = 42{,}086$$

$$c_{64} = \sum_i \sum_j \left[\frac{\left(\sum_s \sum_t X_{1sijt}\right)\left(\sum_s \sum_t X_{2sijt}\right)}{6} \right] = 45{,}181$$

$$d_{34} = \sum_t \left[\frac{\left(\sum_s \sum_i \sum_j X_{2sijt}\right)^2}{18} \right] = 105{,}831$$

$$d_{41} = \sum_s \left[\frac{\left(\sum_i \sum_j \sum_t X_{1sijt}\right)\left(\sum_i \sum_j \sum_t Y_{sijt}\right)}{27} \right] = 18{,}694$$

$$d_{42} = \sum_i \left[\frac{\left(\sum_s \sum_j \sum_t X_{1sijt}\right)\left(\sum_s \sum_j \sum_t Y_{sijt}\right)}{18} \right] = 18{,}741$$

$$d_{43} = \sum_j \left[\frac{\left(\sum_s \sum_i \sum_t X_{1sijt}\right)\left(\sum_s \sum_i \sum_t Y_{sijt}\right)}{18} \right] = 19{,}590$$

$$d_{44} = \sum_t \left[\frac{\left(\sum_s \sum_i \sum_j X_{1sijt}\right)\left(\sum_s \sum_i \sum_j Y_{sijt}\right)}{18} \right] = 18{,}924$$

$$d_{51} = \sum_s \left[\frac{\left(\sum_i \sum_j \sum_t X_{2sijt}\right)\left(\sum_i \sum_j \sum_t Y_{sijt}\right)}{27} \right] = 47{,}097$$

$$d_{52} = \sum_i \left[\frac{\left(\sum_s \sum_j \sum_t X_{2sijt}\right)\left(\sum_s \sum_j \sum_t Y_{sijt}\right)}{18} \right] = 47{,}646$$

$$d_{53} = \sum_j \left[\frac{\left(\sum_s \sum_i \sum_t X_{2sijt}\right)\left(\sum_s \sum_i \sum_t Y_{sijt}\right)}{18} \right] = 50{,}124$$

$$d_{54} = \sum_t \left[\frac{\left(\sum_s \sum_i \sum_j X_{2sijt}\right)\left(\sum_s \sum_i \sum_j Y_{sijt}\right)}{18} \right] = 47{,}596$$

$$d_{61} = \sum_{s}\left[\frac{\left(\sum_{i}\sum_{j}\sum_{t} X_{1sijt}\right)\left(\sum_{i}\sum_{j}\sum_{t} X_{2sijt}\right)}{27}\right] = 41{,}625$$

$$d_{62} = \sum_{i}\left[\frac{\left(\sum_{s}\sum_{j}\sum_{t} X_{1sijt}\right)\left(\sum_{s}\sum_{j}\sum_{t} X_{2sijt}\right)}{18}\right] = 42{,}285$$

$$d_{63} = \sum_{j}\left[\frac{\left(\sum_{s}\sum_{i}\sum_{t} X_{1sijt}\right)\left(\sum_{s}\sum_{i}\sum_{t} X_{2sijt}\right)}{18}\right] = 44{,}346$$

$$d_{64} = \sum_{t}\left[\frac{\left(\sum_{s}\sum_{i}\sum_{j} X_{1sijt}\right)\left(\sum_{s}\sum_{i}\sum_{j} X_{2sijt}\right)}{18}\right] = 42{,}059$$

$$c_{65} = \sum_{i}\sum_{t}\left[\frac{\left(\sum_{s}\sum_{j} X_{1sijt}\right)\left(\sum_{s}\sum_{j} X_{2sijt}\right)}{6}\right] = 42{,}794$$

$$c_{66} = \sum_{j}\sum_{t}\left[\frac{\left(\sum_{s}\sum_{i} X_{1sijt}\right)\left(\sum_{s}\sum_{i} X_{2sijt}\right)}{6}\right] = 44{,}893$$

$$d_{11} = \sum_{s}\left[\frac{\left(\sum_{i}\sum_{j}\sum_{t} Y_{sijt}\right)^{2}}{27}\right] = 21{,}152$$

$$d_{12} = \sum_{i}\left[\frac{\left(\sum_{s}\sum_{j}\sum_{t} Y_{sijt}\right)^{2}}{18}\right] = 21{,}185$$

$$d_{13} = \sum_{j}\left[\frac{\left(\sum_{s}\sum_{i}\sum_{t} Y_{sijt}\right)^{2}}{18}\right] = 22{,}159$$

$$d_{14} = \sum_{t}\left[\frac{\left(\sum_{s}\sum_{i}\sum_{j} Y_{sijt}\right)^{2}}{18}\right] = 21{,}415$$

$$d_{21} = \sum_{s}\left[\frac{\left(\sum_{i}\sum_{j}\sum_{t} X_{1sijt}\right)^{2}}{27}\right] = 16{,}521$$

$$d_{22} = \sum_{i}\left[\frac{\left(\sum_{s}\sum_{j}\sum_{t} X_{1sijt}\right)^{2}}{18}\right] = 16{,}586$$

$$d_{23} = \sum_{j}\left[\frac{\left(\sum_{s}\sum_{i}\sum_{t} X_{1sijt}\right)^{2}}{18}\right] = 17{,}327$$

$$d_{24} = \sum_{t}\left[\frac{\left(\sum_{s}\sum_{i}\sum_{j} X_{1sijt}\right)^{2}}{18}\right] = 16{,}723$$

$$d_{31} = \sum_{s} \left[\frac{\left(\sum_{i} \sum_{j} \sum_{t} X_{2sijt} \right)^2}{27} \right] = 104{,}877$$

$$d_{32} = \sum_{i} \left[\frac{\left(\sum_{s} \sum_{j} \sum_{t} X_{2sijt} \right)^2}{18} \right] = 110{,}645$$

$$d_{33} = \sum_{j} \left[\frac{\left(\sum_{s} \sum_{i} \sum_{t} X_{2sijt} \right)^2}{18} \right] = 114{,}036$$

$$e_1 = \left[\frac{\left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} Y_{sijt} \right)^2}{54} \right] = 21{,}123$$

$$e_2 = \left[\frac{\left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} X_{1sijt} \right)^2}{54} \right] = 16{,}503$$

$$e_3 = \left[\frac{\left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} X_{2sijt} \right)^2}{54} \right] = 104{,}808$$

$$e_4 = \left[\frac{\left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} X_{1sijt} \right) \left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} Y_{sijt} \right)}{54} \right] = 18{,}670$$

$$e_5 = \left[\frac{\left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} X_{2sijt} \right) \left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} Y_{sijt} \right)}{54} \right] = 47{,}051$$

$$e_6 = \left[\frac{\left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} X_{1sijt} \right) \left(\sum_{s} \sum_{i} \sum_{j} \sum_{t} X_{2sijt} \right)}{54} \right] = 41{,}588$$

The application of the method involves the calculation of the sums of squares of the dependent variable and of each of the two independent variables, and the sums of products of each of the independent variables with the dependent variate to be adjusted and with each other. These values are obtained by applying the appropriate formulas in Table 95.

We first test the significance of the interactions. The complete analysis resulting in the tests of significance of the several hypotheses is given in Table 97. Since the adjustment for the two concomitant variates has been obtained from the error term, 2 degrees of freedom ascribed to error have been used in evaluating it. The reduced sum of squares assigned to error is divided by the corresponding number of degrees of freedom to obtain the mean square (1.41) appropriate to testing the significance of the remaining interactions. No significant interaction was found. Therefore, 44 degrees of freedom became available for testing the significance of the main effects.

TABLE 97

TEST OF SIGNIFICANCE OF INTERACTIONS

(Partialing out the effects of both initial score and M.A.)*

Source of variation	D.F.	Σy^2	Σx_1^2	Σx_2^2	$\Sigma x_1 y$	$\Sigma x_2 y$	$\Sigma x_1 x_2$	Adjusted Σy^2	Reduced				
									D.F.	S.S.	M.S.	F.	Hypothesis†
Sex × grade × scholastic × individual (or error)	8	16	26	690	13	21	*−11*	8.47	6	8.47	1.41	....	....
Sex × grade × scholastic	4	25	35	942	29	130	135	1.13	4	0.48	0.12	....	Accepted
Sex × grade × individual	4	18	22	659	19	41	29	3.22	4	2.33	0.58	....	Accepted
Sex × scholastic × individual	4	5	9	515	2	10	19	6.09	4	5.47	1.37	....	Accepted
Grade × scholastic × individual	8	33	28	984	27	112	118	10.11	8	8.92	1.12	....	Accepted
Sex × grade	2	32	53	638	40	138	160	1.49	2	0.62	0.31	....	Accepted
Sex × scholastic	2	2	20	11	6	*−4*	*−14*	.90	2	0.60	0.30	....	Accepted
Sex × individual	2	3	32	119	7	*−14*	*−10*	5.16	2	3.38	1.19	....	Accepted
Grade × scholastic	4	37	29	1299	27	213	138	7.83	4	3.62	0.91	....	Accepted
Grade × individual	4	5	7	445	0	22	38	7.35	4	6.47	1.62	1.15	Accepted
Scholastic × individual	4	10	17	373	9	28	76	6.67	4	6.02	1.51	1.07	Accepted
Coefficients for adjusting Σy^2		1	.2666	.0015	*−1.0328*	*−.0774*	.0400						

* Where $b_1 = .5164$; $b_2 = .0387$.

† The hypothesis tested is a null hypothesis concerning the variation in the same row. For instance, the hypothesis regarding sex × grade × scholastic is that there is no significant interaction between sex, grade, and scholastic standing when the effects of both initial score and mental ages have been partialed out.

TABLE 98
COMPLETE ANALYSIS OF VARIANCE AND COVARIANCE
(Partialing out the effects of both initial score and M.A.)*

Source of variation	D.F.	Σy^2	Σx_1^2	Σx_2^2	$\Sigma x_1 y$	$\Sigma x_2 y$	$\Sigma x_1 x_2$	Adjusted Σy^2	Reduced				
									D.F.	S.S.	M.S.	F	Hypothesis†
Residual.........	46	186	278	6,675	179	697	678	57.21	44	57.21	1.30	.....	....
Sex.............	1	30	19	69	24	45	36	7.70	1	7.21	7.21	5.55	Remains in doubt
Grade...........	2	62	83	5,837	70	594	697	3.25	2	2.89	1.45	1.12	Accepted
Scholastic........	2	1037	824	9,228	920	3073	2757	159.28	2	46.86	23.43	18.02	Rejected
Individual.......	2	292	220	1,022	253	545	471	60.52	2	34.14	17.07	13.13	Rejected
Total	53	1607	1424	22,831	1446	4954	4639	287.96					
Coefficients for adjusting Σy^2		1	.2677	.0027	*−1.0348*	*−.1038*	.0536						

* Where $b_1 = .5174$; $b_2 = .0519$.

† The hypothesis tested is a null hypothesis concerning the variation in the same row. For instance, the hypothesis regarding grade is that there is no significant difference between grade means when the effects of both initial score and mental ages have been partialed out.

TABLE 99

ILLUSTRATION OF TEST OF SIGNIFICANCE WITH REDUCED Σy^2

(Partialing out the effects of X_1 and X_2)

Source of variation	D.F.	Σy^2	Σx_1^2	Σx_2^2	Σx_1y	Σx_2y	Σx_1x_2	Reduced				
								D.F.	S.S.	M.S.	F	Hypothesis
Residual..............	46	186	278	6,675	179	697	678	44	57.21	1.30	....	....
Grade..............	2	62	83	5,837	70	594	697	2	2.89	1.45	1.12	Accepted
Total	48	248	361	12,512	249	1291	1375	46	60.10			

The complete analysis of variance and covariance of the final scores, partialing out the joint effect of initial score and mental-age score, is presented in Table 98. The analysis, which has used all the evidence of the relevant data, led to the conclusion that there was a significant difference among the means of the final scores of the scholastic groups and of the individual orders of development when adjustments were made for the differences in initial and mental-age scores. The difference between the adjusted, means of the sexes was significant at the 5 per cent level.

The whole procedure of making an exact test of significance based on the reduced Σy^2 when there are two independent variates is illustrated for the test of significance for "grade" in Table 99.[7]

PROBLEMS

1. Design an experiment to determine the effect of training upon individual differences.
2. Design a factorial experiment for determining the effect of practice of different levels and kinds upon transfer of training.
3. Design a factorial experiment to determine the effect of various lengths and frequencies of intervals upon learning a fundamental process in arithmetic.
4. Design an educationa experiment which makes use of the Latin-square arrangement.
5. Devise a method of comparing the efficiency from the use of the follow-

[7] For the detailed solution of the problem of estimation, see Ref. 13.

ing three different types of experimental design: Assume the experiment is designed to determine if there is a differential effect of three different treatments (for example, different dietary treatments on school children). Let A, B, and C represent the three treatments; O, the dummy treatment; I, II, III, the three school terms. In Design 1 the possible diet sequences are given by 1, 2, 3, . . . , 24.

DESIGN 1

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
School	I	O	O	O	O	O	O	A	A	A	A	A	A	B	B	B	B	B	B	C	C	C	C	C	A
term	II	A	A	B	B	C	C	O	O	B	B	C	C	O	O	A	A	C	C	O	O	A	A	B	B
	III	B	C	C	A	A	B	B	C	C	O	O	B	A	C	C	O	O	A	A	B	B	O	O	A

DESIGN 2

		1	2	3	4	5	6	7	8
School	I	C	O	A	B	A	C	B	O
term	II	C	O	A	B	A	C	B	O
	III	C	O	A	B	A	C	B	O

In Design 2, the same treatment is administered to the same child throughout the three school terms. The treatments are randomized in blocks of 4 children, who are selected to be as alike as possible.

DESIGN 3

		1	2	3	4	5	6	7	8
School	I	D	B	C	A	O	C	B	A
term	II	B	A	C	O	O	C	A	B
	III	A	C	O	B	A	C	O	B

In Design 3, the treatments within each block of 4 children for each term are rerandomized.

6. Assume there are 15 persons who are to be invited to 35 dinners, and that 3 persons are to take part in each dinner. Arrange the invitations for dinner so that each person is invited 7 times, and 2 persons meet at a dinner just 1 time.

References

1. Alexander, Howard W., "A General Test for Trend," *Psychological Bulletin*, Vol. 43 (1946), pp. 533–557.
2. Anderson, R. L., and Houseman, E. E., "Tables of Orthogonal Polynomial Values Extended to $N = 104$," *Agricultural Experiment Station Bulletin* 297 (1942).
3. Baxter, Brent, "A Study of Reaction Time Using Factorial Design," *Journal of Experimental Psychology*, Vol. 31 (1942), pp. 430–437.
4. Carlson, W. S., and Tinker, Miles A., "Visual Reaction-Time as a Function of Variations in the Stimulus-Figure," *American Journal of Psychology*, Vol. 59 (1946), pp. 450–457.

5. Chapin, F. Stuart, "Some Problems in Field Interviews When Using the Control Group Technique in Studies in the Community," *American Sociological Review*, Vol. VIII (1943), pp. 63–68.

6. Engelhart, Max D., "Suggestions with Respect to Experimentation Under School Conditions," *Journal of Experimental Education*, Vol. 14 (1946), pp. 225–244.

7. Fisher, R. A., *The Design of Experiments*, 2d ed. Edinburgh: Oliver & Boyd, Ltd., 1937.

8. ———, and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research.* Edinburgh: Oliver & Boyd, Ltd., 1943.

9. Garrett, Henry E., and Zubin, Joseph, "The Analysis of Variance in Psychological Research," *Psychological Bulletin*, Vol. 40 (1943), pp. 233–267.

10. Goulden, C. H., *Methods of Statistical Analysis.* New York: John Wiley & Sons, Inc., 1939.

11. Hotelling, Harold, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, Vol. 24 (1933), pp. 417–441; 498–520.

12. ———, "Some Improvements in Weighing and Other Experimental Techniques," *Annals of Mathematical Statistics*, Vol. XV (1944), pp. 297–306.

13. Johnson, Palmer O., and Tsao, Fei, "Factorial Design and Covariance in the Study of Individual Educational Development," *Psychometrika*, Vol. 10 (1945), pp. 133–162.

14. ———, and ———, "Factorial Design in the Determination of Differential Limen Values," *Psychometrika*, Vol. 9 (1944), pp. 107–144.

15. Lindquist, E. F., *Statistical Analysis in Educational Research.* Boston: Houghton Mifflin Company, 1940.

16. Rulon, P. J., "Fisher's t-Test as a Special Case of His z-Test," *Journal of Experimental Education*, Vol. XI (1943), pp. 245–249.

17. Shen, Eugene, "Experimental Design and Statistical Treatment in Educational Research," *Journal of Experimental Education*, VIII (1940), pp. 346–353.

18. Snedecor, G. W., *Statistical Methods*, 4th ed. Ames, Iowa: Iowa Collegiate Press, 1946.

19. Yates, F., "Complex Experiments," *Supplement to Journal of the Royal Statistical Society*, Vol. II, No. 2 (1935), pp. 181–247.

20. ———, "The Design and Analysis of Factorial Experiments," Imperial Bureau of Soil Science, Harpenden, England, *Technical Communication* 35 (1937).

21. ———, "Incomplete Randomized Blocks," *Annals of Eugenics*, Vol. VII, Part VI (1936), pp. 121–140.

22. Wald, Abraham, "On the Efficient Design of Statistical Investigations," *Annals of Mathematical Statistics*, Vol. XIV (1943), pp. 134–140.

CHAPTER XIV

MULTIPLE REGRESSION PROBLEMS

It frequently happens in experimental situations that we are concerned with the problem of estimating or predicting one character from a knowledge of another or of a number of other characters. For prediction or estimation of this kind to be useful, it is necessary that a change in the variable to be predicted is accompanied by some corresponding change in the other variable or variables. Problems of this kind require the quantification of this apparent relationship existing among the variables and are spoken of as *problems in regression.*

In the simple case of the regression of one variate on another, the regression function takes the form

$$Y' = a + b(X - \bar{X}) \tag{14.01}$$

where b is the regression coefficient of Y on X, and Y' is the predicted value of Y for each value of X.

The Multiple Regression Equation. If, instead of having only one independent variate, such as X in the simple case above, we have measures on several independent variables, then we can express the mean value of the dependent variate, Y, in terms of the several independent variates. This is the multivariate case to be treated in this section.

We denote by Y_t the value of the criterion variable, and by X_{it} the value of the ith measurement of the tth individual, respectively. Then the multiple regression equation (or, more accurately, the partial regression equation) for obtaining the simple weighted sum of the measurements, Y'_t may be written

$$Y'_t = a_0 + b_1X_{it} + \cdots + b_kX_{kt} \tag{14.02}$$

where it is assumed that we have the value of the criterion variable, Y_t, and k measurements of each individual. In Equation (14.02), a_0 is a constant; the b's are known as the *partial regression coefficients* and are also constants. Instead of the subscript b_1, for instance, the subscript $Y1.23, \ldots, k$ or $0.123, \ldots, k$ is often used. This subscript indicates more completely than b_1 that the partial regression coefficients show how greatly unit changes in the individual X variables affect Y_t, independently and directly. The values of these constants are to be determined in each case from the available data.

If we let y_t, y'_t, x_1, x_2, . . . , x_k represent the deviations from the respective means of the variables, there is no need for the term a_0 in

Equation (14.02), because

$$\Sigma y_t' = \Sigma x_1 = \Sigma x_2, \cdots, = \Sigma x_k = 0$$

In terms of this notation, Equation (14.02) becomes

$$y_t' = b_1x_1 + b_2x_2 + \cdots + b_kx_k \tag{14.03}$$

In order that y_t' be the best linear estimate of y_t when "best" is considered in the light of the least-squares criterion, $\Sigma(y_t - y_t')$ must be minimized; that is,

$$(y_t - b_1x_1 - b_2x_2 - \cdots - b_kx_k)^2 \tag{14.04}$$

must be minimized. A necessary and sufficient condition for this minimum sum of squares is that the b's satisfy the following system of equations:

$$\left.\begin{array}{l} \Sigma(y - b_1x_1 - b_2x_2 - \cdots - b_kx_k)x_1 = 0 \\ \Sigma(y - b_1x_1 - b_2x_2 - \cdots - b_kx_k)x_2 = 0 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \quad \cdot \\ \Sigma(y - b_1x_1 - b_2x_2 - \cdots - b_kx_k)x_k = 0 \end{array}\right] \tag{14.05}$$

The left members of these equations are the negative of one-half of the partial derivatives of (14.04) with respect to $b_1, b_2, \ldots, b_k$.

Equations (14.05) may be written in the form

$$\left.\begin{array}{l} b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 + \cdots + b_k\Sigma x_1x_k = \Sigma x_1y \\ b_1\Sigma x_1x_2 + b_2\Sigma x_2^2 + \cdots + b_k\Sigma x_2x_k = \Sigma x_2y \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad \cdot \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad \cdot \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad \cdot \\ b_1\Sigma x_1x_k + b_2\Sigma x_kx_2 + \cdots + b_k\Sigma x_kx_k = \Sigma x_ky \end{array}\right] \tag{14.06}$$

The set of equations (14.06) are often called *normal equations*. After computing the necessary sums from the given data and substituting these values in the system of equations, it is possible to solve for the b's in order to obtain the partial regression coefficient. Substituting these values in (14.03), we have the multiple regression equation in deviation form. If, as is usually more convenient, the original measures instead of the deviations are used, these values of the b's may be substituted in Equation (14.02). The value of $a_0 = \bar{Y} - b_1\bar{X}_1 - \cdots - b_k\bar{X}_k$, where the bars denote the mean values of the several variates.

The accuracy with which the regression coefficients or weights enable us to predict or estimate the values of the criterion variable is determined by computing the multiple correlation coefficient. This may be interpreted as the zero order, or total correlation coefficient between the actual values of Y_t and the values Y_t' predicted from the multiple regression equation (14.02). The development of multiple R as a measure of the accuracy of prediction of a multiple regression equation may be observed as follows:

$$\chi^2 = \sum_{t=1}^{N} (Y_t - Y'_t)^2 \qquad (14.07)$$

Let R represent the correlation between the two sets of scores, Y_t and Y'_t; and let $\Sigma y_t^2 = \Sigma(Y_t - \bar{Y}_t)^2 =$ the sum of squares about the mean of Y_t, Equation (14.07) may then be written

$$\chi^2 = \Sigma y_t^2(1 - R^2) \qquad (14.08)$$

from which it follows that the multiple correlation coefficient, R, is the measure of the accuracy with which the criterion scores may be predicted. It may also be pointed out that the multiple correlation is another case of the analysis of variance, that is, of analyzing Σy_t^2 into two parts, one associated with regression and the other a residual.

The value of R, the multiple correlation coefficient, may be readily calculated from the following equation:

$$R^2 = \frac{b_1\Sigma(x_1y) + b_2\Sigma(x_2y) + \cdots + b_k\Sigma(x_ky)}{\Sigma y^2} \qquad (14.09)$$

The normal equations (14.06) may be modified by dividing both members of the first equation by $\sqrt{\Sigma x_1^2 \cdot \Sigma y^2}$; both members of the second equation by $\sqrt{\Sigma x_2^2 \cdot \Sigma y^2}$; and . . . of the kth equation by $\sqrt{\Sigma x_k^2 \cdot \Sigma y^2}$.

This modification yields the following system:

$$\left.\begin{array}{l} \beta_1 + \beta_2 r_{12} + \cdots + \beta_{1k} = r_{1Y} \\ \beta_1 r_{12} + \beta_2 + \cdots + \beta_{2k} = r_{2Y} \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ \cdot \quad \cdot \quad \cdot \quad \cdot \\ \beta_2 r_{1k} + \beta_k + \cdots + \beta_{kk} = r_{kY} \end{array}\right] \qquad (14.10)$$

where $\beta_1 = b_1\sqrt{\frac{\Sigma x_1^2}{\Sigma y^2}}$; $\beta_2 = b_2\sqrt{\frac{\Sigma x_2^2}{\Sigma y^2}}$, $\cdots$; and $\beta_k = b_k\sqrt{\frac{\Sigma x_k^2}{\Sigma y^2}}$. The β's are known as the *standard partial regression coefficients*, to distinguish them from the b's, the partial regression coefficients. The β's are the partial regression coefficients for the variates expressed in standard measure form, thus rendering them independent of the original units of measurement and giving measures of the comparative weight attributable to each of the independent variates. In terms of the β's, the multiple correlation coefficient is given as

$$R^2_{Y\cdot 123,\cdots,k} = \beta_1 r_{1Y} + \beta_2 r_{2Y} + \cdots + \beta_k r_{kY} \qquad (14.11)$$

A systematic procedure often used for the solution of the system of normal equations (14.06) or (14.10) is known as the *Doolittle method,* after its formulator, an engineer with the United States Coast and Geodetic Survey. Doolittle, in 1878, introduced a method which was due to various improvements over Gauss's method of solving simultaneous linear equations by direct substitutions. Some modifications of Doolittle's method have occurred from time to time, but the essential features of his method persist (Refs. 8 and 9). This method is applied below, but first it is desirable to enumerate what is involved in the complete analysis of a multiple regression problem.

We have described above the method of setting up the multiple regression equation and of calculating the criterion of its predictive accuracy, the multiple correlation coefficient. The values of the b's or the β's alone, however, give a very incomplete description of the relationships between the dependent variable, Y, and the independent variates, $X_1, \ldots, X_k$. They do not indicate whether all—or, if not all, which—of the independent variates are significantly related to the dependent variate; nor can the confidence intervals or fiducial limits be specified from them within which the true values of the regression coefficients are to be found. The standard error of the sum or the difference between two regression coefficients may be needed. Where no apparent relation is found between the dependent variate and one or more of the independent variates, it is often desirable to omit such variates from the regression equation. It may also at times be desirable to add one or more new independent variates to the original battery. Occasionally, there may be an interest in the multiple correlation between a certain set of independent variates and each of several dependent variates. Finally, when predictions for each individual have been made from the multiple regression equation, we are interested in the accuracy of each individual prediction and in setting up a confidence interval for each individual.

To facilitate the carrying out of most of the above analysis, Fisher (Ref. 13) has suggested the use of a set of auxiliary quantities, C_{pq} $(p, q = 1, 2, \cdots, k)$. The quantities $C_{p1}, C_{p2}, \ldots, C_{pk}$ are the solutions of the set of equations (14.06) with the right-hand side of the pth equation replaced by 1, and of the other equations by 0. The relations between the regression coefficients and the auxiliaries, C's, are given by

$$b_i = \sum_{q=1}^{k} C_{iq} \sum (X_{qv}) \quad (i = 1, 2, \cdots, k) \qquad (14.12)$$

For example, for the case of 3 independent variates the 3 systems of equations are obtained by using for the right members of the equations

1, 0, 0 for the first system; 0, 1, 0 for the second system; and 0, 0, 1 for the third system:

$$\left.\begin{array}{l}A_1\Sigma x_1^2 + A_2\Sigma x_1x_2 + A_3\Sigma x_1x_3 = 1 \quad 0 \quad 0\\ A_1\Sigma x_1x_2 + A_2\Sigma x_2^2 + A_3\Sigma x_2x_3 = 0 \quad 1 \quad 0\\ A_1\Sigma x_1x_3 + A_2\Sigma x_2x_3 + A_3\Sigma x_3^2 = 0 \quad 0 \quad 1\end{array}\right] \qquad (14.13)$$

The three solutions for these three sets of equations may be written

$$\left.\begin{array}{l}A_1 = C_{11}, C_{12}, C_{13}\\ A_2 = C_{12}, C_{22}, C_{23}\\ A_3 = C_{13}, C_{23}, C_{33}\end{array}\right] \qquad (14.14)$$

Once the 6 values of C are known, then the partial regression coefficients may be obtained in any particular case by calculating Σx_1y, Σx_2y, Σx_3y and substituting in the following formulas:

$$\left.\begin{array}{l}b_1 = C_{11}\Sigma x_1y + C_{12}\Sigma x_2y + C_{13}\Sigma x_3y\\ b_2 = C_{12}\Sigma x_1y + C_{22}\Sigma x_2y + C_{23}\Sigma x_3y\\ b_3 = C_{13}\Sigma x_1y + C_{23}\Sigma x_2y + C_{33}\Sigma x_3y\end{array}\right] \qquad (14.15)$$

Problem XIV.1. The complete analysis of a regression problem. We shall illustrate the complete analysis of a regression problem as it was carried out in a study of predicting in the School of Agriculture in the University of Minnesota. In this problem it was of interest to secure the correlation coefficients between the several variates. Furthermore, the use of correlation coefficients in the normal equations provides the same order of magnitude for all the quantities at any given step in the solution. Their use is also advantageous in the use of the check column to be described later. The standard partial regression coefficients rather than the partial regression coefficients are used because of the interest in comparing the relative importance of the independent variates, which originally were in different units of measurement. For this case the auxiliary set of quantities, the C's used for securing the b's have been supplanted by what we call the g's for securing the β's.

We have observed 213 individuals with 1 dependent variable and 5 independent variables. Let us denote the dependent variable or the criterion by Y, and the independent variables by X_1, X_2, X_3, X_4, and X_5. The scores observed are as follows:

Y: honor-point ratios

X_1: age

X_2: Iowa Silent Reading Test score

X_3: Otis raw score

X_4: previous education in years

X_5: School of Agriculture Reading Test total score

We wish to predict the honor-point ratio from the measures of the independent variates. The following steps are pursued:

Step 1. Compute all the intercorrelations and the standard deviations. Let us define:

$$\bar{X}_i = \frac{\Sigma X_i}{N};\ \bar{Y} = \frac{\Sigma Y}{N};\ x_i = X_i - \bar{X}_i;\ y = Y - \bar{Y}$$

$$s_i^2 = \frac{\Sigma x_i^2}{N-1};\ s_y^2 = \frac{\Sigma y^2}{N-1};\ r_{ij} = \frac{\Sigma x_i x_j}{N s_i s_j};\ r_{iy} = \frac{\Sigma x_i y}{N s_i s_y}$$

where $i \neq j$ and $i, j = 1, \cdots, 5$. All the measures in our case are summarized as follows:

$N = 213; \bar{X}_1 = 15.9296; \bar{X}_2 = 161.9061; \bar{X}_3 = 37.8498; \bar{X}_4 = 8.9531$
$\bar{X}_5 = 90.0235;\ \bar{Y} = 2.3362;\ s_1^2 = 5.34245869;\ s_2^2 = 246.33860141$
$s_3^2 = 109.11357981;\ s_4^2 = 5.26540141;\ s_5^2 = 587.14968357$
$s_y^2 = .70191238;\ s_1s_2 = 36.27745774;\ s_1s_3 = 24.14404430$
$s_1s_4 = 5.30378969;\ s_1s_5 = 56.00734941;\ s_2s_3 = 163.94782712$
$s_2s_4 = 36.01487742;\ s_2s_5 = 380.31255769;\ s_3s_4 = 23.96928698$
$s_3s_5 = 253.11264376;\ s_4s_5 = 55.60196191;\ r_{12} = .0300$
$r_{13} = .0983;\ r_{14} = .1100;\ r_{15} = .0470;\ r_{23} = .7143;\ r_{24} = .0960$
$r_{25} = .8203;\ r_{34} = .1821;\ r_{35} = .7230;\ r_{45} = .1124;\ r_{1y} = .1784$
$r_{2y} = .6505;\ r_{3y} = .5164;\ r_{4y} = .0993;\ r_{5y} = .6704$[1]

Step 2. Compute Fisher's auxiliary statistics (g_{ij})'s. The 5 systems of simultaneous equations to be solved are

$$\left.\begin{array}{lccccc}
 & \multicolumn{5}{c}{\text{Right members of system}} \\
 & (1) & (2) & (3) & (4) & (5) \\
g_1 + r_{12}g_2 + r_{13}g_3 + r_{14}g_4 + r_{15}g_5 = & 1 & 0 & 0 & 0 & 0 \\
r_{12}g_1 + \quad g_2 + r_{23}g_3 + r_{24}g_4 + r_{25}g_5 = & 0 & 1 & 0 & 0 & 0 \\
r_{13}g_1 + r_{23}g_2 + \quad g_3 + r_{34}g_4 + r_{35}g_5 = & 0 & 0 & 1 & 0 & 0 \\
r_{14}g_1 + r_{24}g_2 + r_{34}g_3 + \quad g_4 + r_{45}g_5 = & 0 & 0 & 0 & 1 & 0 \\
r_{15}g_1 + r_{25}g_2 + r_{35}g_3 + r_{45}g_4 + \quad g_5 = & 0 & 0 & 0 & 0 & 1
\end{array}\right] \qquad (14.16)$$

The values obtained for the g's in the first system will be designated by g_{11}, g_{21}, g_{31}, g_{41}, and g_{51}. The values obtained for the g's in the second, the third, the fourth, and the fifth systems will be designated by g_{12}, g_{22}, g_{32}, g_{42}, and g_{52}; by g_{13}, g_{23}, g_{33}, g_{43}, and g_{53}; by g_{14}, g_{24}, g_{34}, g_{44}, and by g_{54}; by g_{15}, g_{25}, g_{35}, g_{45}, and g_{55}, respectively. It is worthy of note that

$$g_{ij} = g_{ji} \quad (i \neq j;\ i, j = 1, \cdots, 5) \qquad (14.17)$$

[1] We have used 4 decimal places in our calculations. This is likely a minimum number with the number of equations and of unknowns used. As the number of equations and unknowns increases, the Doolittle and other similar methods of elimination require increasingly larger numbers of decimal places. For example, probably at least 10 places would be necessary for 10 unknowns if a final answer of 1-place accuracy is wanted (Ref. 17).

TABLE 100

SYSTEMATIC PROCEDURE IN SOLVING THE SYSTEM OF EQUATIONS (14.16)

Directions	Multiplier	A	B	C	D	E	F	F'	F''	F'''	$F^{(4)}$	Check
(1): Eq. (I)		$1.0000g_1$	$.0300g_2$	$.0983g_3$	$.1100g_4$	$.0470g_5$	= 1	0	0	0	0	2.2853
(2): Eq. (II)		$.0300g_1$	$1.0000g_2$	$.7143g_3$	$.0960g_4$	$.8203g_5$	= 0	1	0	0	0	3.6606
(3): (I) × (−.0300)	$-B_1$	$-.0300g_1$	$-.0009g_2$	$-.0030g_3$	$-.0033g_4$	$-.0014g_5$	= −.0300	0	0	0	0	−.0686
(4): (2) + (3)			$.9991g_2$	$.7113g_3$	$.0927g_4$	$.8189g_5$	= −.0300	1.0000	0	0	0	3.5920
(5): (4) ÷ .9991			$1.0000g_2$	$.7119g_3$	$.0928g_4$	$.8196g_5$	= −.0300	1.0009	0	0	0	3.5952
(6): Eq. (III)		$.0983g_1$	$.7143g_2$	$1.0000g_3$	$.1821g_4$	$.7230g_5$	= 0	0	1	0	0	3.7177
(7): (I) × (−.0983)	$-C_1$	$-.0983g_1$	$-.0030g_2$	$-.0097g_3$	$-.0108g_4$	$-.0046g_5$	= −.0983	0	0	0	0	−.2247
(8): (4) × (−.7119)	$-C_5$		$-.7113g_2$	$-.5064g_3$	$-.0660g_4$	$-.5830g_5$	= .0214	−.7119	0	0	0	−2.5572
(9): (6) + (7) + (8)				$.4839g_3$	$.1053g_4$	$.1354g_5$	= −.0769	−.7119	1.0000	0	0	.9358
(10): (9) ÷ .4839				$1.0000g_3$	$.2176g_4$	$.2798g_5$	= −.1589	−1.4712	2.0665	0	0	1.9338
(11): Eq. (IV)		$.1100g_1$	$.0960g_2$	$.1821g_3$	$1.0000g_4$	$.1124g_5$	= 0	0	0	1	0	2.5005
(12): (I) × (−.1100)	$-D_1$	$-.1100g_1$	$-.0033g_2$	$-.0108g_3$	$-.0121g_4$	$-.0052g_5$	= −.1100	0	0	0	0	−.2514
(13): (4) × (−.0928)	$-D_5$		$-.0927g_2$	$-.0660g_3$	$-.0086g_4$	$-.0760g_5$	= .0028	−.0928	0	0	0	−.3333
(14): (9) × (−.2176)	$-D_{10}$			$-.1053g_3$	$-.0229g_4$	$-.0295g_5$	= .0167	.1549	−.2176	0	0	−.2037
(15): (11) + (12) + (13) + (14)					$.9564g_4$	$.0017g_5$	= −.0905	.0621	−.2176	1.0000	0	1.7121
(16): (15) ÷ .9564					$1.0000g_4$	$.0018g_5$	= −.0946	.0649	−.2275	1.0456	0	1.7902
(17): Eq. (V)		$.0470g_1$	$.8203g_2$	$.7230g_3$	$.1124g_4$	$1.0000g_5$	= 0	0	0	0	1	3.7027
(18): (I) × (−.0470)	$-E_1$	$-.0470g_1$	$-.0014g_2$	$-.0046g_3$	$-.0052g_4$	$-.0022g_5$	= −.0470	0	0	0	0	−.1074
(19): (4) × (−.8196)	$-E_5$		$-.8189g_2$	$-.5830g_3$	$-.0760g_4$	$-.6712g_5$	= .0246	−.8196	0	0	0	−2.9441
(20): (9) × (−.2798)	$-E_{10}$			$-.1354g_3$	$-.0295g_4$	$-.0379g_5$	= .0215	.1992	−.2798	0	0	−.2619
(21): (15) × (−.0018)	$-E_{16}$				$-.0017g_4$	$.0000g_5$	= .0002	−.0001	.0004	−.0018	0	−.0030
(22): (17)+(18)+(19)+(20)+(21)						$.2887g_5$	= −.0007	−.6205	−.2794	−.0018	1.0000	.3863
(23): (22) ÷ .2887						$1.0000g_5$	= −.0024	−2.1493	−.9678	−.0062	3.4638	1.3381

For our problem, (14.16) becomes

$$\left.\begin{array}{lrrrrr}
\text{(I)}: 1.0000g_1 + .0300g_2 + .0983g_3 + .1100g_4 + .0470g_5 & & & & & \\
 & = 1 & 0 & 0 & 0 & 0 \\
\text{(II)}: .0300g_1 + 1.0000g_2 + .7143g_3 + .0960g_4 + .8203g_5 & & & & & \\
 & = 0 & 1 & 0 & 0 & 0 \\
\text{(III)}: .0983g_1 + .7143g_2 + 1.0000g_3 + .1821g_4 + .7230g_5 & & & & & \\
 & = 0 & 0 & 1 & 0 & 0 \\
\text{(IV)}: .1100g_1 + .0960g_2 + .1821g_3 + 1.0000g_4 + .1124g_5 & & & & & \\
 & = 0 & 0 & 0 & 1 & 0 \\
\text{(V)}: .0470g_1 + .8203g_2 + .7230g_3 + .1124g_4 + 1.0000g_3 & & & & & \\
 & = 0 & 0 & 0 & 0 & 1
\end{array}\right]$$

A systematic procedure often used for the solution of such a system of equations is shown in Table 100. A convenient check column is often carried along, to the right of these computations. The first and second entries of this check column are found by adding all other entries in their respective rows. The third entry is found in two ways, thus yielding a check on the accuracy of the arithmetical computations. The first way consists in the addition of all other entries in the third row. The other way consists in operating on the first entry in accordance with the directions given at the left. The other entries in the check column are found in a similar way.[2]

The values of g_{51}, g_{52}, g_{53}, g_{54}, and g_{55} can be read directly from the last row, numbered (23):

$$g_{51} = -.0024; \qquad g_{52} = -2.1493; \qquad g_{53} = -.9678; \qquad g_{54} = -.0062;$$
$$g_{55} = 3.4638$$

We get g_{41}, g_{42}, g_{43}, g_{44}, and g_{45} as follows:

Substitute g_{51} in Eq. (16, Table 100) and use column F in the right-hand member:

$$g_{41} + .0018(-.0024) = -.0946$$
$$g_{41} = -.0946$$

Substitute g_{52} in Eq. (16, Table 100) and use column F' in the right member:

$$g_{42} + .0018(-2.1493) = .0649$$
$$g_{42} = .0688$$

Substitute g_{53} in Eq. (16, Table 100) and use column F'' in the right member:

$$g_{43} + .0018(-.9678) = -.2275$$
$$g_{43} = -.2258$$

[2] It should be noted that errors occurring in the rounding of the original correlations are not accounted for by the check column.

Substitute g_{54} in Eq. (16, Table 100) and use column F''' in the right member:

$$g_{44} + .0018(-.0062) = 1.0456$$
$$g_{44} = 1.0456$$

Substitute g_{55} in Eq. (16, Table 100) and use column $F^{(4)}$ in the right member:

$$g_{45} + .0018(3.4638) = 0$$
$$g_{45} = -.0062$$

To obtain g_{31}, g_{32}, g_{33}, g_{34}, and g_{35}:

Substitute g_{41} and g_{51} in Eq. (10, Table 100) and use F in the right member:

$$g_{31} + .2176(-.0946) + .2798(-.0024) = -.1589$$
$$g_{31} = -.1377$$

Substitute g_{42} and g_{52} in Eq. (10, Table 100) and use F' in the right member:

$$g_{32} + .2176(.0688) + .2798(-2.1493) = -1.4712$$
$$g_{32} = -.8847$$

Substitute g_{43} and g_{53} in Eq. (10, Table 100) and use F'' in the right member:

$$g_{33} + .2176(-.2258) + .2798(-.9678) = 2.0665$$
$$g_{33} = 2.3864$$

Substitute g_{44} and g_{54} in Eq. (10, Table 100) and use F''' in the right member:

$$g_{34} + .2176(1.0456) + .2798(-.0062) = 0$$
$$g_{34} = -.2258$$

Substitute g_{45} and g_{55} in Eq. (10, Table 100) and use $F^{(4)}$ in the right member:

$$g_{35} + .2176(-.0062) + .2798(3.4638) = 0$$
$$g_{35} = -.9678$$

To obtain g_{21}, g_{22}, g_{23}, g_{24}, and g_{25}:

Substitute g_{31}, g_{41}, and g_{51} in Eq. (5, Table 100) and use F in the right member:

$$g_{21} + .7119(-.1377) + .0928(-.0946) + .8196(-.0024) = -.0300$$
$$g_{21} = .0788$$

Substitute g_{32}, g_{42}, and g_{52} in Eq. (5, Table 100) and use F' in the right member:

$$g_{22} + .7119(-.8847) + .0928(.0688) + .8196(-2.1493) = 1.0009$$
$$g_{22} = 3.3859$$

Substitute g_{33}, g_{43}, and g_{53} in Eq. (5, Table 100) and use F'' in the right member:

$$g_{23} + .7119(2.3864) + .0928(-.2258) + .8196(-.9678) = 0$$
$$g_{23} = -.8847$$

Substitute g_{34}, g_{44}, and g_{54} in Eq. (5, Table 100) and use F''' in the right member:

$$g_{24} + .7119(-.2258) + .0928(1.0456) + .8196(-.0062) = 0$$
$$g_{24} = .0688$$

Substitute g_{35}, g_{45}, and g_{55} in Eq. (5, Table 100) and use $F^{(4)}$ in the right member:

$$g_{25} + .7119(-.9678) + .0928(-.0062) + .8196(3.4638) = 0$$
$$g_{25} = 2.1493$$

To obtain g_{11}, g_{12}, g_{13}, g_{14}, and g_{15}:

Substitute g_{21}, g_{31}, g_{41}, and g_{51} in Eq. (1, Table 100) and use F in the right member:

$$g_{11} + .0300(.0788) + .0983(-.1377) + .1100(-.0946) + .0470(-.0024) = 1$$
$$g_{11} = 1.0217$$

Substitute g_{22}, g_{32}, g_{42}, and g_{52} in Eq. (1, Table 100) and use F' in the right member:

$$g_{12} + .0300(3.3859) + .0983(-.8847) + .1100(.0688) + .0470(-2.1493) = 0$$
$$g_{12} = .0788$$

Substitute g_{23}, g_{33}, g_{43}, and g_{53} in Eq. (1, Table 100) and use F'' in the right member:

$$g_{13} + .0300(-.8847) + .0983(2.3864) + .1100(-.2258) + .0470(-.9678) = 0$$
$$g_{13} = -.1377$$

Substitute g_{24}, g_{34}, g_{44}, and g_{54} in Eq. (1, Table 100) and use F''' in the right member:

$$g_{14} + .0300(.0688) + .0983(-.2258) + .1100(1.0456) + .0470(-.0062) = 0$$
$$g_{14} = -.0946$$

Substitute g_{25}, g_{35}, g_{45}, and g_{55} in Eq. (1, Table 100) and use $F^{(4)}$ in the right member:

$$g_{15} + .0300(-2.1493) + .0983(-.9678) + .1100(-.0062) + .0470(3.4638) = 0$$
$$g_{15} = -.0024$$

The accuracy of the (g_{ij})'s $(i \neq j)$ can be checked by the equation $g_{ij} = g_{ji}$; and the accuracy of the (g_{ii})'s can be checked by a method illustrated by Wallace and Snedecor (Ref. 25).[3] It is shown that to obtain g_{11}, the sum of products of the last two members (regardless of sign) in each section in column (F) is found; similarly, for g_{22} the same procedure is followed in column (F'), and so on. In our problem

$$\begin{aligned}
g_{11} &= 1 + .0300(.0300) + .0769(.1589) + .0905(.0946) + .0007(.0024) \\
&\qquad = 1.0217 \\
g_{22} &= 1.0000(1.0009) + .7119(1.4712) + .0621(.0649) + .6205(2.1493) \\
&\qquad = 3.3859 \\
g_{33} &= 1.0000(2.0665) + .2176(.2275) + .2794(.9678) = 2.3864 \\
g_{44} &= 1.0000(1.0456) + .0018(.0062) = 1.0456 \\
g_{55} &= 1.0000(3.4638) = 3.4638
\end{aligned}$$

Since we have checked all our results, we shall summarize them as follows:

$$\begin{array}{llll}
g_{11} = 1.0217; & g_{12} = .0788; & g_{13} = -.1377; & g_{14} = -.0946; \\
 & & & \quad g_{15} = -.0024 \\
g_{21} = .0788; & g_{22} = 3.3859; & g_{23} = -.8847; & g_{24} = .0688; \\
 & & & \quad g_{25} = 2.1493 \\
g_{31} = -.1377; & g_{32} = -.8847; & g_{33} = 2.3864; & g_{34} = -.2258; \\
 & & & \quad g_{35} = -.9678 \\
g_{41} = -.0946; & g_{42} = .0688; & g_{43} = -.2258; & g_{44} = 1.0456; \\
 & & & \quad g_{45} = -.0062 \\
g_{51} = -.0024; & g_{52} = 2.1493; & g_{53} = -.9678; & g_{54} = -.0062; \\
 & & & \quad g_{55} = 3.4638
\end{array}$$

Step 3. Compute $R_{y.12345}$, the multiple correlation between Y and X_1, X_2, X_3, X_4, X_5.

Define:

$$\beta_i = \sum_j g_{ij} r_{jy} \qquad (i, j = 1, \cdots, 5) \qquad (14.18)$$

where β_i is the standard partial regression coefficient. For our problem, we have

$$\begin{aligned}
\beta_1 &= g_{11}r_{1y} + g_{12}r_{2y} + g_{13}r_{3y} + g_{14}r_{4y} + g_{15}r_{5y} = .1514 \\
\beta_2 &= g_{21}r_{1y} + g_{22}r_{2y} + g_{23}r_{3y} + g_{24}r_{4y} + g_{25}r_{5y} = .3256 \\
\beta_3 &= g_{31}r_{1y} + g_{32}r_{2y} + g_{33}r_{3y} - g_{34}r_{4y} + g_{35}r_{5y} = -.0390 \\
\beta_4 &= g_{41}r_{1y} + g_{42}r_{2y} + g_{43}r_{3y} + g_{44}r_{4y} + g_{45}r_{5y} = .0109 \\
\beta_5 &= g_{51}r_{1y} + g_{52}r_{2y} + g_{53}r_{3y} + g_{54}r_{4y} + g_{55}r_{5y} = .4232
\end{aligned}$$

Define again:

$$R^2_{y.12345} = \sum_i \beta_i r_{iy} \qquad (i = 1, \cdots, 5) \qquad (14.19)$$

[3] This method does not provide a complete guarantee of accuracy, since in some instances large errors in the solution might give only small deviations of the left from the right member of the equation, and since some deviations are to be expected when only a limited number of decimal places are carried along.

For our problem, we have

$$R^2_{y.12345} = \beta_1 r_{1y} + \beta_2 r_{2y} + \beta_3 r_{3y} + \beta_4 r_{4y} + \beta_5 r_{5y} = .50346861$$

Therefore, we obtain

$$R_{y.12345} = .7096$$

Step 4. Test the significance of $R_{y.12345}$. This can be done through the use of the variance ratio. The method is shown below.

ANALYSIS-OF-VARIANCE TABLE

Source of variation	Sum of squares	D.F.	Mean square
Not associated with regression	$(1 - R^2)\Sigma y^2$	$N - m - 1$	$\dfrac{(1 - R^2)\Sigma y^2}{N - m - 1}$
Associated with regression	$R^2\Sigma y^2$	m	$\dfrac{R^2\Sigma y^2}{m}$
Total	Σy^2	$N - 1$	

$$F \text{ (variance ratio)} = \frac{R^2(N - m - 1)}{m(1 - R^2)} \tag{14.20}$$

For our problem,

$$F = \frac{.5035(207)}{5(.4965)} = 41.98$$

Referring to the F tables (Table IV, Appendix) with $n_1 = 5$ and $n_2 = 207$, we have $P < .01$. Therefore, we conclude that the value of the multiple correlation is significantly different from zero.

Step 5. Test the significance of (β_i)'s.

Define:

$$s_{\beta_i} = \sqrt{\frac{(1 - R^2_{y.12345})g_{ii}}{N - m - 1}} \quad (i = 1, \cdots, 5) \tag{14.21}$$

where s_{β_i} is the standard error of β_i. For our problem we have:

$$s_{\beta_1} = \sqrt{\frac{(1 - R^2_{y.12345})g_{11}}{N - m - 1}} = .0495$$

$$s_{\beta_2} = \sqrt{\frac{(1 - R^2_{y.12345})g_{22}}{N - m - 1}} = .0901$$

$$s_{\beta_3} = \sqrt{\frac{(1 - R^2_{y.12345})g_{33}}{N - m - 1}} = .0757$$

$$s_{\beta_4} = \sqrt{\frac{(1 - R^2_{y.12345})g_{44}}{N - m - 1}} = .0501$$

$$s_{\beta_5} = \sqrt{\frac{(1 - R^2_{y.12345})g_{55}}{N - m - 1}} = .0911$$

The test of significance of each β_i is given by

$$t_{\beta_i} = \frac{\beta_i}{s_{\beta_i}} \tag{14.22}$$

with $N - m - 1$ degrees of freedom. For our problem, we obtain

$$t_{\beta_1} = 3.059; \quad t_{\beta_2} = 3.614; \quad t_{\theta_3} = -.515$$
$$t_{\beta_4} = .218; \quad t_{\beta_5} = 4.645$$

Referring to the t-table with 207 degrees of freedom, we find that β_1, β_2, and β_5 are significant at the 1 per cent level, and that β_3 and β_4 are not significantly different from zero. Therefore, we can omit the independent variables X_3 and X_4.[4]

Step 6. The omission of X_3: Let us denote by β'_i and g'_{ij} the new standard regression coefficient and auxiliary statistics, respectively. By mathematical derivations, we have

$$\beta'_i = \beta_i - \frac{g_{i3}}{g_{33}}\beta_3 \quad (i = 1, 2, 4, 5) \tag{14.23}$$

$$g'_{ij} = g_{ij} - \frac{g_{i3}g_{j3}}{g_{33}} \quad (i, j = 1, 2, 4, 5) \tag{14.24}$$

For our problem, we can easily obtain

$$\beta'_1 = \beta_1 - \frac{g_{13}}{g_{33}}\beta_3 = .1492; \qquad \beta'_2 = \beta_2 - \frac{g_{23}}{g_{33}}\beta_3 = .3112$$

$$\beta'_4 = \beta_4 - \frac{g_{43}}{g_{33}}\beta_3 = .0072; \qquad \beta'_5 = \beta_5 - \frac{g_{53}}{g_{33}}\beta_3 = .4074$$

$$g'_{11} = g_{11} - \frac{g_{13}^2}{g_{33}} = 1.0138; \qquad g'_{12} = g_{12} - \frac{g_{13}g_{23}}{g_{33}} = .0278$$

$$g'_{14} = g_{14} - \frac{g_{13}g_{43}}{g_{33}} = -.1076; \qquad g'_{15} = g_{15} - \frac{g_{13}g_{53}}{g_{33}} = -.0582$$

$$g'_{22} = g_{22} - \frac{g_{23}^2}{g_{33}} = 3.0579; \qquad g'_{24} = g_{24} - \frac{g_{23}g_{43}}{g_{33}} = -.0149$$

$$g'_{25} = g_{25} - \frac{g_{23}g_{53}}{g_{33}} = -2.5081; \qquad g'_{44} = g_{44} - \frac{g_{43}^2}{g_{33}} = 1.0242$$

$$g'_{45} = g_{45} - \frac{g_{43}g_{53}}{g_{33}} = -.0978; \qquad g'_{55} = g_{55} - \frac{g_{53}^2}{g_{33}} = 3.0713$$

Proceeding as before, we have

$$R^2_{y.1245} = \beta'_1 r_{1y} + \beta'_2 r_{2y} + \beta'_4 r_{4y} + \beta'_5 r_{5y} = .5028880$$
$$R_{y.1245} = .7091$$
$$F = \frac{R^2(N - m' - 1)}{m'(1 - R^2)} = 52.6067$$

where $m' = 4$. Referring to the F-table with $n_1 = 4$ and $n_2 = 208$, we find that $P < .01$. Therefore, we conclude that the R is significant.

[4] For a discussion of "suppression" variables which might increase the multiple correlation even if they correlate zero or near zero with the criterion, see Refs. 14 and 27.

In testing the significance of (β'_i)'s, we obtain

$$s_{\beta'_1} = \sqrt{\frac{(1 - R^2_{y.1245})g'_{11}}{N - m' - 1}} = .0492$$

$$s_{\beta'_2} = \sqrt{\frac{(1 - R^2_{y.1245})g'_{22}}{N - m' - 1}} = .0855$$

$$s_{\beta'_4} = \sqrt{\frac{(1 - R^2_{y.1245})g'_{44}}{N - m' - 1}} = .0495$$

$$s_{\beta'_5} = \sqrt{\frac{(1 - R^2_{y.1245})g'_{55}}{N - m' - 1}} = .0857$$

Consequently, we obtain

$$t_{\beta'_1} = \frac{\beta'_1}{s_{\beta'_1}} = 3.033; \qquad t_{\beta'_2} = \frac{\beta'_2}{s_{\beta'_2}} = 3.640$$

$$t_{\beta'_4} = \frac{\beta'_1}{s_{\beta'_4}} = .145; \qquad t_{\beta'_5} = \frac{\beta'_5}{s_{\beta'_5}} = 4.754$$

Referring to the t-table with 208 degrees of freedom, we find that β'_1, β'_2, and β'_5 are significant at the 1 per cent level and that β_4 is not significantly different from 0. Therefore, it is desirable to omit the independent variable X_4.

Step 7. The omission of both[5] X_3 and X_4. Let us denote by β''_i and g''_{ij} the new standard regression coefficients and auxiliary statistics, respectively. By mathematical derivations, we have

$$\beta''_i = \beta'_i - \frac{g'_{i4}}{g'_{44}}\beta'_4 \qquad (i = 1, 2, 5) \qquad (14.25)$$

$$g''_{ij} = g'_{ij} = \frac{g_{i4}g_{j4}}{g_{44}} \qquad (i, j = 1, 2, 5) \qquad (14.26)$$

For our problem, we can easily obtain

$$\beta''_1 = \beta'_1 - \frac{g'_{14}}{g'_{44}}\beta'_4; \qquad \beta''_2 = \beta'_2 - \frac{g'_{24}}{g'_{44}}\beta'_4$$
$$= .1500 \qquad\qquad = .3113$$

$$\beta''_5 = \beta'_5 - \frac{g'_{54}}{g'_{44}}\beta'_4$$
$$= .4081$$

$$g''_{11} = g'_{11} - \frac{g'^2_{14}}{g'_{44}} = 1.0025; \qquad g''_{12} = g'_{12} - \frac{g'_{14}g'_{24}}{g'_{44}} = .0262$$

$$g''_{15} = g'_{15} - \frac{g'_{14}g'_{45}}{g'_{44}} = -.0685; \qquad g''_{22} = g'_{22} - \frac{g'^2_{24}}{g'_{44}} = 3.0577$$

$$g''_{25} = g'_{25} - \frac{g'_{24}g'_{45}}{g'_{44}} = -2.5095; \qquad g''_{55} = g'_{55} - \frac{g'^2_{45}}{g'_{44}} = 3.0620$$

[5] The advantage of the use of the g-statistics over the method of resolving the normal equations depends upon the number of independent variates. If the number of originally independent variates is 6 or more, or perhaps 5, and if 2 are to be eliminated, the use of the g-statistics is advisable.

Proceeding as before, we have

$$\begin{aligned} R^2_{y.125} &= \beta''_1 r_{1y} + \beta''_2 r_{2y} + \beta''_5 r_{5y} = .50285089 \\ R_{y.125} &= .7091 \\ F &= \frac{R^2(N - m'' - 1)}{m''(1 - R^2)} = 70.480 \end{aligned}$$

where $m'' = 3$. Referring to F-table with $n_1 = 3$ and $n_2 = 209$, we have $P < .01$. Therefore, we conclude that the R is significantly different from zero.

In testing the significance of the standard partial regression coefficients, we have

$$\begin{aligned} s_{\beta''_1} &= \sqrt{\frac{(1 - R^2_{y.125})g''_{11}}{N - m'' - 1}} = .0488 \\ s_{\beta''_2} &= \sqrt{\frac{(1 - R^2_{y.125})g''_{22}}{N - m'' - 1}} = .0853 \\ s_{\beta''_5} &= \sqrt{\frac{(1 - R^2_{y.125})g''_{55}}{N - m'' - 1}} = .0853 \end{aligned}$$

Consequently, we have

$$t_{\beta''_1} = \frac{\beta''_1}{s_{\beta''_1}} = 3.074; \qquad t_{\beta''_2} = \frac{\beta''_2}{s_{\beta''_2}} = 3.649; \qquad t_{\beta''_5} = \frac{\beta''_5}{s_{\beta''_5}} = 4.784$$

Referring to the t-table with 209 degrees of freedom, we find that all the standard partial regression coefficients are significant. Therefore, we conclude that the three independent variables X_1, X_2, and X_5 should be used in order to predict the dependent variable Y.

Step 8. Set up the formula for the prediction of Y from X_1, X_2, and X_5. First we calculate the partial regression coefficients. Define:

$$b_i = \beta_i \frac{s_y}{s_i} \quad (i = 1, \cdots, 5) \tag{14.27}$$

Similarly,

$$b'_i = \beta'_i \frac{s_y}{s_i} \quad (i = 1, 2, 4, 5) \tag{14.28}$$

$$b''_i = \beta''_i \frac{s_y}{s_i} \quad (i = 1, 2, 5) \tag{14.29}$$

For our problem, we do not need to use b_i and b'_i. Therefore, we have

$$b''_1 = \beta''_1 \frac{s_y}{s_1} = .0544; \qquad b''_2 = \beta''_2 \frac{s_y}{s_2} = .0166; \qquad b''_5 = \beta''_5 \frac{s_y}{s_5} = .0141$$

where $s_y = .837802$, $s_1 = 2.311376$, $s_2 = 15.695178$, and $s_5 = 24.231172$ which are calculated from the results in Step 1. Denoting by $\hat{Y}$ the predicted Y-score, we have

$$\hat{Y} = \bar{Y} + \sum_i b_i x_i \quad (i = 1, 2, \cdots, 5) \tag{14.30}$$

Similarly,

$$\hat{Y} = \bar{Y} + \sum_i b_i x_i \qquad (i = 1, 2, 4, 5) \qquad (14.31)$$

$$\hat{Y} = \bar{Y} + \sum_i b_i'' x_i \qquad (i = 1, 2, 5) \qquad (14.32)$$

Again, for our problem we simply use (14.32). Then we obtain

$$\begin{aligned}\hat{Y} &= \bar{Y} + b_1''(X_1 - \bar{X}_1) + b_2''(X_2 - \bar{X}_2) + b_5''(X_5 - \bar{X}_5)\\ &= 2.3362 - .0544(15.9296) - .0166(161.9061)\\ &\quad - .0141(90.0235) + .0544X_1 + .0166X_2 + .0141X_5\\ &= .0544X_1 + .0166X_2 + .0141X_5 - 2.4873\end{aligned}$$

It is to be noted that this predicted score refers to the true mean Y-score of all the individuals who have the same specific scores of X_1, X_2, and X_5 in the population. In other words, in the long run, the true mean Y-score of all the individuals who have identical scores for X_1, X_2, and X_5 in the population will approach $\hat{Y}$ within a fiducial limit which we may set up. Since in our example we do not find two individuals with the same scores for X_1, X_2, and X_5, it is difficult to verify the accuracy of the predicted score.

Step 9. Compute the standard error of an individual predicted score. The general formula using the auxiliary statistics (g_{ij})'s for predicting the standard error of an individual predicted score is

$$s_{\hat{y}} = \sqrt{\frac{s_y^2(1 - R^2_{y.123\,\cdots\,m})}{N - m - 1}\left[1 + \sum_i \frac{g_{ii}}{s_i^2} x_i^2 + 2 \sum_{\substack{i,j\\ i<j}} \frac{g_{ij}}{s_i s_j} x_i x_j\right]} \qquad (14.33)$$

where $i, j = 1, \cdots, m$; m is the number of independent variables; s_y^2, the estimate of the population variance σ_y^2; and the s_i's and s_j's and the x's are defined as before.[6] For our problem, we simply use $i, j = 1, 2, 5$ and change m to $m'' = 3$ and (g)'s to (g'')'s. Thus we have

$$\begin{aligned}s_{\hat{y}} &= \sqrt{\frac{s_y^2(1 - R^2_{y.125})}{N - m'' - 1}\left[1 + \frac{g_{11}''}{s_1^2} x_1^2 + \frac{g_{22}''}{s_2^2} x_2^2 + \frac{g_{55}''}{s_5^2} x_5^2 + \frac{2g_{12}''}{s_1 s_2} x_1 x_2 + \frac{2g_{15}''}{s_1 s_5} x_1 x_5 + \frac{2g_{25}''}{s_2 s_5} x_2 x_5\right]}\\ &= \sqrt{.001670[1 + .1876x_1^2 + .0124x_2^2 + .0052x_5^2 + .0014x_1x_2 - .0024x_1x_5 - .0132x_2x_5]}\end{aligned}$$

[6] The working formula can also be written as follows:

$$s_{\hat{y}} = \sqrt{\frac{y^2(1 - R^2{}_{y.123\,\cdots\,m})}{N - m - 1}\left[\frac{1}{N} + \sum_i \frac{g_{ii}}{\Sigma x_i{}^2} + 2 \sum_{\substack{i,j\\ i<j}} \frac{g_{ii}}{\sqrt{\Sigma x_i{}^2 \, \Sigma x_j{}^2}} x_i x_j\right]}$$

Step 10. Find the fiducial limits.

$$\text{Fiducial limits } (p) = \hat{Y} \pm t_{(q)}(s_{\hat{y}})$$

where $p + q = 1$ and the t-value is obtained by referring to the t-table with $N - m - 1$ (for our problem, $N - m'' - 1$) degrees of freedom. It may be stated with a confidence coefficient of $100p$ per cent that the true mean Y-score of all the individuals who have identical X_1, X_2, and X_5-scores will lie in the range of $\hat{Y} \pm t_{(q)}(s_{\hat{y}})$. It is customary to make $p = .99$ or $.95$. For our problem, $N - m'' - 1 = 209$. Referring to the t-table with 209 degrees of freedom, we find that $t_{.01} = 2.600$ and $t_{.05} = 1.972$. Therefore, we have

$$\begin{aligned} \text{Fiducial limits } (.99) &= \hat{Y} \pm 2.600(s_{\hat{y}}) \\ \text{Fiducial limits } (.95) &= \hat{Y} \pm 1.972(s_{\hat{y}}) \end{aligned}$$

Step 11. Practical application: To find the fiducial limits for the true mean Y-score for the following individual values:

$$X_1 = 16; \qquad X_2 = 163; \qquad X_5 = 80; \qquad Y = 2.316$$

$$\hat{Y} = -2.4873 + .0544(16) + .0166(163) + .0141(80) = 2.217$$

$$x_1 = .0704; \qquad x_2 = 1.0939; \qquad x_5 = -10.0235$$

$$s_{\hat{y}} = \sqrt{.001670 \begin{bmatrix} 1 + .1876(.0704)^2 + .0124(1.0939)^2 \\ + .0052(-10.0235)^2 + .0014(.0704)(1.0939) - .0024 \\ (.0704)(-10.0235) - .0132(1.0939)(-10.0235) \end{bmatrix}}$$

$$= .0530$$

$$\begin{aligned} \text{Fiducial limits } (.99) &= 2.217 \pm 2.600(.0530) = (2.079, 2.355) \\ \text{Fiducial limits } (.95) &= 2.217 \pm 1.972(.0530) = (2.112, 2.322) \end{aligned}$$

The Discriminant Function. The ordering of things into classes is a basic procedure of empirical science. In fact, the rigorousness of the basis of scientific classification is an index of the development of a field as a science. Statistical methods are available which can be profitably applied to the problem of discriminating between different populations and classifying them. The aspect of the problem to be discussed here deals with the statistical uses of multi-measurement for differentiating between two or more groups of individuals, things, or events. This is frequently a problem in economics, education, psychology, or in the various fields of science. For instance, individuals upon whom several measurements are available are to be classified into groups with a minimum of overlapping. The traditional method is to compute the significance of the difference between the means of groups taking each character separately. This method is inefficient in that it does not make possible the evaluation of the relative amount of information for differentiation provided by the several measurements; neither does it combine the information taking into account the interrelations, if they exist, between the characters dealt with. From this observation, the problem is clearly one analogous to multiple regression; that is, a weighted sum of the measurements is needed as in multiple regression. The difference lies in the nature of the criterion which is, in the problem discussed here, qualitative rather than quantitative as in the case of multiple regression. That is, the dependent variable is a dichotomy or a multiple classification. The particular statistic for the solution of this problem, which is called the *discriminant function*, was developed by Fisher (Ref. 10). The essential property of this function, which is a linear function of the observations, is that it will distinguish better than any other linear function between the specified groups on whom common measurements are available. The principle upon which the discriminant function rests is that the linear functions of the measurements will maximize the ratio of the difference between the specific means to the standard deviations within classes. This type of problem is also closely related to that studied by Hotelling (Refs. 15 and 16) resulting in his generalization of "Student's" ratio, or Hotelling's T, as it is usually called, which is a powerful tool for testing the significance between mean values of different multivariate normal populations under the assumptions of equal variances and equal covariances. Closely related also is the statistic developed by Mahalanobis (Ref. 19) and studied further by Bose and Roy (Ref. 3), leading to the studentized form of the distribution, in statistics called the *generalized distance function*, D^2. By the use of D^2, different multivariate populations can be not merely discriminated but also classified, that is, D^2 contributes both to the problem of testing significance and of estimation. The treatment here is limited to the discriminant function.

We first present the formulation and solution of the mathematical problem. Then a practical application is presented.

Two Groups. If we have samples of N_1 and N_2 observations, respectively, and make p measurements $X_1, \ldots, X_p$ on each individual, consider first the question: What linear function of the measurements will maximize the ratio of the difference between the means of the two classes to the standard deviation within classes? The linear function is represented by

$$\alpha = \sum_i \lambda_i x_i \quad (i = 1, \cdots, p) \qquad (14.34)$$

Let the difference between means of x_i be represented by d_i, where $i = 1, \cdots, p$ for the p measurements. Represent the sum of squares or products from the specific means within classes by S_{ij}, where $i, j = 1, \cdots, p$. Then for any linear function, α, of the measurements, the difference between the means of α in the two specific groups is

$$D = \sum_i \lambda_i d_i \quad (i = 1, \cdots, p) \qquad (14.35)$$

while the variance of α within classes is proportional to

$$S_a = \sum_i \sum_j \lambda_i \lambda_j S_{ij} \quad (i, j = 1, \cdots, p) \qquad (14.36)$$

The particular function which best discriminates the two groups will be one for which the ratio D^2/S_a is greatest, by variation of the p coefficients, $\lambda_1, \ldots, \lambda_p$, independently. Mathematically, we should seek the solution for each λ:

$$\frac{\partial}{\partial \lambda}\left(\frac{D^2}{S}\right) = 0 \qquad (14.37)$$

which reduces to

$$\frac{D}{S^2}\left(2S\frac{\partial D}{\partial \lambda} - D\frac{\partial S}{\partial \lambda}\right) = 0 \qquad (14.38)$$

and consequently,

$$\frac{1}{2}\frac{\partial S}{\partial \lambda} = \frac{S}{D} \cdot \frac{\partial D}{\partial \lambda} \qquad (14.39)$$

where it may be noticed that S/D is a factor common to the p unknown λ's. Therefore, the coefficients required are proportional to the solutions of the normal equations:

$$\left.\begin{array}{l} S_{11}\lambda_1 + \cdots + S_{1p}\lambda_p = d_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ S_{p1}\lambda_1 + \cdots + S_{pp}\lambda_p = d_p \end{array}\right] \qquad (14.40)$$

Let us define:

$$L_i = \sqrt{S_{ii}}\,\lambda_i \quad (i = 1, \cdots, p) \qquad (14.41)$$

In (14.40) we divide the ith equation by $\sqrt{S_{ii}}$, where $i = 1, \cdots, p$. Then we have the following set of normal equations:

$$\left.\begin{array}{l} r_{11}L_1 + \cdots + r_{1p}L_p = \dfrac{d_1}{\sqrt{S_{11}}} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ r_{p1}L_1 + \cdots + r_{pp}L_p = \dfrac{d_p}{\sqrt{S_{pp}}} \end{array}\right] \qquad (14.42)$$

We can easily solve (14.42) for L's by Fisher's method of auxiliary statistics, in which unity is substituted for each of the $d_i/\sqrt{S_{ii}}$'s in turn, while the others are made equal to zero as follows:

$$\left.\begin{array}{l} r_{11}L_1 + \cdots + r_{1p}L_p = 1, \cdots, 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ r_{p1}L_1 + \cdots + r_{pp}L_p = 0, \cdots, 1 \end{array}\right] \qquad (14.43)$$

Let us define the means of α for these two groups:

$$\bar{\alpha}_1 = \sum_i \lambda_i \bar{X}_{1i} \quad (i = 1, \cdots, p) \tag{14.44}$$

$$\bar{\alpha}_2 = \sum_i \lambda_i \bar{X}_{2i} \quad (i = 1, \cdots, p) \tag{14.45}$$

when $\bar{X}_{1i}$ is the mean value of X_i for the first group and $\bar{X}_{2i}$ is the mean value of X_i for the second group. We wish to test the hypothesis:

$$H_0: E(\bar{\alpha}_1) = E(\bar{\alpha}_2) \quad \left(\begin{array}{c}\textit{E is the notation for the expectation}\\ \textit{of a parameter}\end{array}\right) \tag{14.46}$$

that is, the hypothesis that there is no significant difference between two groups for the function α. By mathematical deductions, the sums of squares due to "within groups" and "between groups" are

$$\text{"Within groups"} \quad D \text{ with } n_2 = N_1 + N_2 - p - 1 \tag{14.47}$$

$$\text{"Between"} \quad \frac{N_1 N_2}{N_1 + N_2} D^2 \text{ with } n_1 = p \tag{14.48}$$

Then the test of H_0 is given by

$$F = \frac{N_1 + N_2 - p - 1}{p} \cdot \frac{N_1 N_2}{N_1 + N_2} D$$

If we reject the hypothesis, H_0, we may conclude that the obtained values of λ's are the assigned weights of the measurements which best discriminate these two groups. Then the next problem arises such that if we have another individual to be observed by making the same measurements, $X_1, \ldots, X_p$, on him, we wish to know to which group he belongs. Wald (Ref. 26) has shown two methods for solving this problem: (1) when N_1 and N_2 are sufficiently large, and (2) when N_1 and N_2 are small. For the time being, we assume that N_1 and N_2 are sufficiently large. By using Wald's criterion, let us denote by π_1 and π_2 the populations of the first group and the second group, respectively. The hypothesis tested in this problem is that the individual is drawn from π_1. First we calculate:

$$\bar{\alpha}_1 = \sum_i \sum_j S_{ij} \bar{X}_{1i} d_j = \lambda_1 \bar{X}_{11} + \cdots + \lambda_p \bar{X}_{1p} \qquad (i, j = 1, \cdots, p) \tag{14.49}$$

$$\bar{\alpha}_2 = \sum_i \sum_j S_{ij} \bar{X}_{2i} d_j = \lambda_1 \bar{X}_{21} + \cdots + \lambda_p \bar{X}_{2p} \qquad (i, j = 1, \cdots, p) \tag{14.50}$$

$$U = \sum_i \sum_j S_{ij} X_i d_j = \lambda_1 X_1 + \cdots + \lambda_p X_p \qquad (i, j = 1, \cdots, p) \tag{14.51}$$

where $\bar{\alpha}_1$, $\bar{\alpha}_2$, $\bar{X}_{1i}$, $\bar{X}_{2i}$, S_{ij}, and d_j are defined as before, X_i is the value obtained by this individual on the ith measurement; and U is the value obtained by the individual for the linear function α. Then the critical region for rejecting the hypothesis with the least risk of both kinds of error, that is, accepting the hypothesis when it is false and rejecting the hypothesis when it is true, is given by

$$U \geqq \frac{\bar{\alpha}_1 + \bar{\alpha}_2}{2} \qquad (14.52)$$

Problem XIV.2. Discrimination between two groups. There were two classes in the College of Science, Literature and Arts in the University of Minnesota. One class was taking the course Physics 7, which was more advanced than Physics 1 taken by the other class. Three measurements were available for each individual: mathematical test score, American Council Examination (A.C.E.) test score, and honor-point ratio (H.P.R.). Let us denote the mathematical test score by X_1, the A.C.E. test score by X_2, and the H.P.R. by X_3. The calculated measures are summarized in Table 101.

TABLE 101
CALCULATED MEASURES FOR TWO GROUPS

	Physics 1	Physics 7
N	111	257
ΣX_1	9,728	23,746
$\bar{X}_1$	87.6396	92.3969
ΣX_2	3,450	14,411
$\bar{X}_2$	31.0811	56.0739
ΣX_3	128.6	326.1
$\bar{X}_3$	1.1586	1.2689
ΣX_1^2	905,694	2,388,412
ΣX_2^2	118,846	823,945
ΣX_3^2	200.84	534.17
ΣX_1X_2	307,220	1,349,410
ΣX_1X_3	11,756.0	31,974.6
ΣX_2X_3	4,240.8	19,122.3
d_1	4.7573	
d_2	24.9928	
d_3	0.1103	

In Table 102 are recorded the computations leading to the pooled sum of squares and products within the two groups. In the line of totals, the entries are the sum of squares and products of the entire 368 individuals.

In the line for groups are put down the sums of squares and products of the group sums in Table 101, calculated in the manner characteristic of analysis of variance and covariance. As an example, the entry for

TABLE 102

CALCULATION OF THE CORRELATION COEFFICIENTS AND THE STANDARD DEVIATIONS WITHIN TWO GROUPS

	X_1		X_2		X_3	
X_1	Total..............	3,294,106		1,656,630		43,730.6
	Groups............	3,046,614.8969		1,633,888.2977		41,401.0826
	Within groups S_{11} =	247,491.1031	S_{12} =	22,741.7023	S_{13} =	2,329.5174
	$\sqrt{S_{11}}$ =	497.4847	$\sqrt{S_{11}}\,\sqrt{S_{22}}$ =	82,468.2875	$\sqrt{S_{11}}\,\sqrt{S_{33}}$ =	6,528.9892
	s_1 =	26.003942	r_{12} =	0.275763	r_{13} =	0.356796
X_2	Total............................			942,791		23,363.1
	Groups..........................			915,311.1344		22,282.7356
	Within groups.....................		S_{22} =	27,479.8656	S_{23} =	1,080.3644
			$\sqrt{S_{22}}$ =	165.7705	$\sqrt{S_{22}}\,\sqrt{S_{33}}$ =	2,175.5720
			s_2 =	8.664963	r_{23} =	0.496589
X_3			Total............................			735.01
			Groups..........................			562.7696
			Within groups...................		S_{33} =	172.2404
					$\sqrt{S_{33}}$ =	13.1240
					s_3 =	0.686002

column X_1 in row X_1 of Table 102 is

$$\frac{(9728)^2}{111} + \frac{(23,746)^2}{257} = 3,046,614.8969$$

and for column X_2, row X_1,

$$\frac{(9728)(3450)}{111} + \frac{(23,746)(14,411)}{257} = 1,633,888.2977$$

The differences in the third line are the sums of squares and products of deviations within the group. The calculation of the standard deviations and the correlations now proceed in the usual manner.

As examples,

$$s_1 = \frac{\sqrt{247,491.1031}}{\sqrt{366}} = 26.003942$$

$$r_{12} = \frac{22,741.7023}{(497.4847)(165.7705)} = .275763$$

The degrees of freedom used, 366, are those within the two groups, $(N_1 - 1) + (N_2 - 1)$.

Calculate:

$$\frac{d_1}{\sqrt{S_{11}}} = .009563; \qquad \frac{d_2}{\sqrt{S_{22}}} = .150767; \qquad \frac{d_3}{\sqrt{S_{33}}} = .008404$$

Consequently, we obtain the following set of normal equations:

$$\begin{aligned}
1.000000L_1 + .275763L_2 + .356796L_3 &= .009563\\
.275763L_1 + 1.000000L_2 + .496589L_3 &= .150767\\
.356796L_1 + .496589L_2 + 1.000000L_3 &= .008404
\end{aligned}$$

The solutions of L_1, L_2, and L_3 are carried out in Table 103.

In Table 103 a convenient check column is often carried along, to the right of these computations. The first and second entries of this check column are found by adding all other entries in their respective rows. The third entry is found in two ways, thus yielding a check on the accuracy of the arithmetical computations. The first way consists of addition of all entries in the third row; the other way consists of operating on the first entry in the check column in accordance with the directions given at the left. The other entries in the check column are found in a similar way.

The values of k_{31}, k_{32}, and k_{33} can be read directly[7] from the last row, (10) in Table 103:

$$k_{31} = -.339403; \qquad k_{32} = -.614720; \qquad k_{33} = 1.426361$$

[7] The k-values are used in the calculations of the L's as noted on page 351.

TABLE 103
SOLUTIONS OF L'S AND λ'S FOR TWO GROUPS

Directions	Multiplier	(A)	(B)	(C)	(D)	(D')	(D'')	Check
(1): Eq. I		1.000000	.275763	.356796	= 1	0	0	2.632559
(2): Eq. II		.275763	1.000000	.496589	= 0	1	0	2.772352
(3): Eq. I · (−.275763)	$-B_1$	− *.275763*	− *.076045*	− *.098391*	=− *.275763*	0	0	− *.725962*
(4): (2) + (3)			.923955	.398198	=− *.275763*	1.000000	0	2.046390
(5): (4) ÷ (.923955)			1.000000	.430971	=− *.298459*	1.082304	0	2.214816
(6): Eq. III		.356796	.496589	1.000000	= 0	0	1	2.853385
(7): Eq. I · (−.356796)	$-C_1$	− *.356796*	− *.098391*	− *.127303*	=− *.356796*	0	0	− *.939286*
(8): (4) · (−.430971)	$-C_5$		− *.398198*	− *.171612*	= .118846	− *.430971*	0	− *.881935*
(9): (6) + (7) + (8)				.701085	=− *.237950*	− *.430971*	1.000000	1.032164
(10): (9) ÷ (.701085)				1.000000	=− *.339403*	− *.614720*	1.426361	1.472238

k_{11} = 1.163065	k_{12} = − *.152186*	k_{13} = − *.339403*
k_{21} = − *.152186*	k_{22} = 1.347230	k_{23} = − *.614720*
k_{31} = − *.339403*	k_{32} = − *.614720*	k_{33} = 1.426361
L_1 = − *.014675*	L_2 = .196496	L_3 = − *.083938*
λ_1 = − *.00002950*	λ_2 = .00118535	λ_3 = − *.00639576*

In order to obtain k_{21}, k_{22}, and k_{23}:

Substitute k_{31} in Eq. (5), Table 103, using column (D) in the right member:

$$k_{21} + .430971(-.339403) = -.298459$$
$$k_{21} = -.152186$$

Substitute k_{32} in Eq. (5), Table 103, using (D′) in the right member:

$$k_{22} + .430971(-.614720) = 1.082304$$
$$k_{22} = 1.347230$$

Substitute k_{33} in Eq. (5), Table 103, using (D″) in the right member:

$$k_{23} + .430971(1.426361) = 0$$
$$k_{23} = -.614720$$

To obtain k_{11}, k_{12}, and k_{13}:

Substitute k_{21} and k_{31} in Eq. (1), Table 103, using (D):

$$k_{11} + .275763(-.152186) + .356796(-.339403) = 1$$
$$k_{11} = 1.163065$$

Substitute k_{22} and k_{32} in Eq. (1), Table 103, using (D′):

$$k_{12} + .275763(1.347230) + .356796(-.614720) = 0$$
$$k_{12} = -.152186$$

Substitute k_{23} and k_{33} in Eq. (1), Table 103, using (D″):

$$k_{13} + .275763(-.614720) + .356796(1.426361) = 0$$
$$k_{13} = -.339403$$

It is noted that

$$k_{ij} = k_{ji} \qquad (i \neq j,\ i, j = 1, 2, 3)$$

This is a good check on the calculation of k_{ij} $(i \neq j)$. The check of k_{ii} $(i = 1, 2, 3)$ can be carried out easily. To obtain k_{11}, the sum of products of the last two numbers (regardless of sign) in each section in column (D) is found. For k_{22} do the same in column (D′), and so on. We have

$$k_{11} = 1.000000 + .275763(.298459) + .237950(.339403) = 1.163065$$
$$k_{22} = 1.000000(1.082304) + .430971(.614720) = 1.347230$$
$$k_{33} = 1.000000(1.426361) = 1.426361$$

The values of L_1, L_2, and L_3 are obtained by calculating the following equations:

$$L_1 = \frac{d_1}{\sqrt{S_{11}}} k_{11} + \frac{d_2}{\sqrt{S_{22}}} k_{12} + \frac{d_3}{\sqrt{S_{33}}} k_{13}$$
$$L_2 = \frac{d_1}{\sqrt{S_{11}}} k_{21} + \frac{d_2}{\sqrt{S_{22}}} k_{22} + \frac{d_3}{\sqrt{S_{33}}} k_{23}$$
$$L_3 = \frac{d_1}{\sqrt{S_{11}}} k_{31} + \frac{d_2}{\sqrt{S_{22}}} k_{32} + \frac{d_3}{\sqrt{S_{33}}} k_{33}$$

Consequently, the values of λ_1, λ_2, and λ_3 are obtained by calculating the following equations:

$$\lambda_1 = \frac{L_1}{\sqrt{S_{11}}}; \qquad \lambda_2 = \frac{L_2}{\sqrt{S_{22}}}; \qquad \lambda_3 = \frac{L_3}{\sqrt{S_{33}}}$$

All these values are shown in Table 103.

The next step is to calculate the value of D. As a check we can use two equations:

$$D = L_1 \frac{d_1}{\sqrt{S_{11}}} + L_2 \frac{d_2}{\sqrt{S_{22}}} + L_3 \frac{d_3}{\sqrt{S_{33}}}$$

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3$$

In our case: $D = .028779$

This value is also the "within" sum of squares. The "between" sum of squares is

$$\frac{N_1 N_2}{N_1 + N_2} \cdot D^2 = .064186$$

The test of significance between two groups on the variable α is given in Table 104.

TABLE 104
ANALYSIS OF VARIANCE OF α BETWEEN AND WITHIN GROUPS

Source of variation	D.F.	S.S.	M.S.	F	Hypothesis
Within groups	364	.028779	.00007906	....	
Between groups	3	.064186	.021395	270.617	Rejected
Total	367	.092965			

Referring to the F-table with $n_1 = 3$ and $n_2 = 364$, we find $p < .01$. Therefore, we reject the hypothesis of homogeneous groups; and the relative value of the variable α for discriminating between groups is apparently indicated by the weights of the different measurements:

$$\lambda_1 = -.00002950; \qquad \lambda_2 = .00118535; \qquad \lambda_3 = -.00639576$$

Now suppose an individual is given these same measurements and obtains

$$X_1 = 80; \qquad X_2 = 40; \qquad X_3 = 1.5$$

We wish to know to which group this individual should be assigned. First, we calculate

$$\bar{\alpha}_1 = -.00002950(87.6396) + .00118535(31.0811) - .00639576(1.1586) = .026846$$

$$\bar{\alpha}_2 = -.00002950(92.3969) + .00118535(56.0739) - .00639576(1.2689) = .055626$$

$$U = -.00002950(80) + .00118535(40) - .00639576(1.5) = .035460$$

$$\frac{\bar{\alpha}_1 + \bar{\alpha}_2}{2} = .041236$$

It is evident that $$U < \frac{\bar{\alpha}_1 + \bar{\alpha}_2}{2}$$

Therefore, we may conclude that this individual should be assigned to the class Physics 1.

PROBLEMS

1. What methods of multivariate analysis other than those reported in this chapter are available? Which of these are applicable to problems in the field of your interest? [In this connection, see Tintner, Gerhard, "Some Applications of Multivariate Analyses to Economic Data," *Journal of the American Statistical Association*, Vol. 41, pp. 472–500 (December, 1946.)]
2. Specify the problem of factor analysis in psychology as a special application of the theory of regression. [See Holzinger, K. J., and Harmon, H. H., *Factor Analysis* (University of Chicago Press, 1941); Thompson, G. H., *The Factorial Analysis of Human Ability:* (Houghton Mifflin Company, 2d. ed., 1946); Thurstone, L. L., *Multiple-Factor Analysis: A Development and Expansion of the Vectors of Mind* (University of Chicago Press, 1947.)]
3. The following data for a random sample of 50 students were taken from a study dealing with the prediction of achievement of freshmen in a particular college of the University of Minnesota:

   Y_1 = honor-point ratio at the end of the fall quarter
   Y_2 = honor-point ratio at the end of the freshman year
   X_1 = score on Johnson Science Application Test
   X_2 = score on an English test
   X_3 = score on the Cooperative Algebra Test
   X_4 = percentile rank in high-school graduation class transformed to probits

   In this problem you are to do the following:

   (a) Set up the multiple regression equation for predicting either Y_1 or Y_2 from X_1, X_2, X_3, and X_4.
   (b) Test the significance of the
      (1) Standard partial regression coefficients (the betas).
      (2) Multiple correlation coefficient.
      (3) Differences between the respective betas.
   (c) Set up a new multiple regression equation eliminating the independent variable or variables that are not statistically significant.
   (d) Repeat (b).
   (e) Calculate the standard error of the predicted score and set up the confidence interval, with a confidence coefficient of 98 per cent for Students 8, 25, 43, and 47.

Student No.	Y_1 Honor-point ratio f(1)	Y_2 Honor-point ratio fws(2)	X_1 Johnson Science (3)	X_2 Coop. English (4)	X_3 Algebra (5)	X_4 High-school P.R. converted into S.D. units (6)
1	1.65	1.57	56	80	46	4.72
2	1.29	1.38	34	113	48	5.64
3	.88	1.15	32	94	75	4.92
4	1.29	.11	55	47	32	6.48
5	.94	.54	37	126	59	5.05
6	.80	.83	32	81	7	4.53
7	.46	.50	33	115	34	5.95
8	1.00	.85	62	148	58	6.65
9	.72	.06	28	84	36	4.39
10	.31	.76	41	119	16	4.33
11	.13	.64	20	69	7	5.03
12	.20	.39	47	77	24	4.77
13	.44	.70	33	106	16	4.48
14	1.00	1.49	44	80	31	4.69
15	.21	.60	41	84	28	5.05
16	1.27	1.67	28	79	15	4.67
17	1.06	1.65	47	109	64	5.10
18	.71	.84	50	92	23	4.23
19	− .07	− .24	31	93	20	3.44
20	1.65	1.43	43	74	32	5.81
21	1.59	.50	59	87	58	4.87
22	− .12	.43	38	95	14	4.01
23	.27	.35	29	72	38	3.77
24	.12	.41	27	106	18	5.10
25	.00	.41	38	71	22	5.00
26	1.12	1.12	40	122	12	5.39
27	.29	.37	41	84	26	4.77
28	1.00	1.12	46	123	32	4.29
29	1.31	.98	55	111	24	5.95
30	1.56	1.14	52	86	15	4.87
31	1.71	1.08	46	76	17	5.71
32	.13	.33	48	111	23	4.95
33	.53	1.06	59	105	61	5.81
34	.12	.60	25	98	45	4.50
35	.29	.69	42	72	0	5.67
36	.75	.58	39	115	10	5.99
37	.09	.17	37	116	40	4.75
38	.73	.85	28	49	21	5.81
39	.09	.17	37	116	40	4.75
40	1.41	.84	62	89	50	4.69
41	.24	.11	24	58	13	5.00
42	.24	.48	32	117	17	4.77
43	2.00	1.56	47	72	15	5.25
44	.07	.18	45	115	24	4.64
45	1.00	1.30	52	86	44	5.15
46	1.57	1.67	65	147	84	6.08
47	− .62	− .77	31	98	36	2.67
48	.77	.21	54	129	64	4.12
49	1.47	1.33	51	131	5	5.84
50	.56	0.00	30	75	21	4.12

4. The following statistics were derived from data collected in a study dealing with the relation between instruction in a course in college biology and the students' belief in the efficacy of certain commercial preparations and home remedies. The criterion was the score on the test, Y. The independent variates, X_1, X_2, X_3, X_4, and X_5 are specified below. In this problem:

   (a) Set up the multiple regression equation for estimating Y from X_1, X_2, X_3, X_4, and X_5.
   (b) Test the significance of the
      (1) Standard partial regression coefficients.
      (2) Multiple correlation coefficient.
   (c) Of the variance of the dependent variate Y accounted for by the combined effect of the independent variates, calculate the proportion assignable to each of the independent variates. (See Johnson, Palmer O., "The Differential Function of Examinations," *Journal of Educational Research*, Vol. 30 (1936), pp. 93–103.)
   (d) Test the significance of the difference between the two largest partial regression coefficients.
   (e) Find the 5 per cent fiducial limits for the largest partial regression coefficient.
   (f) Calculate the partial correlation coefficient, $r_{YX_1.X_5}$, and test its significance.

Zero order correlations: $N = 223$

$r_{12} = .452$				
$r_{13} = .303$	$r_{23} = .638$			
$r_{14} = .324$	$r_{24} = .274$	$r_{34} = .171$		
$r_{15} = .147$	$r_{25} = .326$	$r_{35} = .190$	$r_{45} = .189$	
$r_{1y} = .514$	$r_{2y} = .621$	$r_{3y} = .542$	$r_{4y} = .197$	$r_{5y} = .134$

$s_1 = 6.50$	$\bar{X}_1 = 22.52$	Where Y = score on application in hygiene
$s_2 = 23.78$	$X_2 = 80.54$	X_1 = score on test of facts and principles in hygiene
$s_3 = 4.20$	$\bar{X}_3 = 23.4$	X_2 = score on vocabulary test in hygiene
$s_4 = 0.956$	$\bar{X}_4 = 4.95$	X_3 = score on final examination in hygiene
$s_5 = 1.00$	$\bar{X}_5 = 5.60$	X_4 = transformed high-school percentile ranks
$s_y = 4.89$	$\bar{Y} = 32.08$	X_5 = transformed College Aptitude Test percentile ranks

5. The following data were collected on two groups of students in an experimental investigation of the relative efficacy of two different

methods in teaching agricultural chemistry at the high-school level Compare the two groups with respect to the set of multiple measurements made at the beginning of the experiment.

Topical Assignment Group

Pupil	X_1	X_2	X_3	X_4
1	110	1.37	38	27
2	81	1.62	18	3
3	111	2.49	26	10
4	110	1.96	20	7
5	95	0.86	15	8
6	85	0.56	14	10
7	97	1.38	25	9
8	90	0.25	13	3
9	85	0.51	21	11
10	83	0.78	21	5
11	83	1.15	22	7
12	100	2.24	31	15
13	106	0.72	22	3
14	92	1.36	20	6
15	94	1.25	16	11
16	96	0.62	12	9
17	83	0.90	16	8
18	113	2.65	28	8
19	104	1.61	21	11
20	93	1.51	20	10

Discussion Group

Pupil	X_1	X_2	X_3	X_4
1	103	1.50	26	15
2	115	1.74	28	12
3	104	1.36	25	7
4	85	0.25	20	3
5	84	0.53	17	5
6	87	0.25	23	6
7	93	2.00	34	6
8	112	1.34	24	4
9	123	2.64	44	16
10	106	0.75	20	10
11	99	2.11	24	13
12	80	0.45	22	7
13	112	1.96	40	16
14	91	1.19	17	5
15	77	0.42	14	6
16	96	1.68	20	11
17	85	0.90	15	2
18	115	1.65	22	7
19	117	1.75	26	11

X_1 = Intelligence quotient based on Kuhlman-Anderson Tests.

X_2 = Honor-point ratio of previous year's work.

X_3 = Score on pretest of knowledge of facts and principles examination administered at the beginning of the term.

X_4 = Score on pretest of Glenn-Welton Chemistry Achievement Test administered at the beginning of the year.

References

1. Bartlett, M. S., "The Standard Errors of Discriminant Functions," *Supplement to Journal of the Royal Statistical Society,* Vol. VI (1939), pp. 169–173.
2. Beall, Geoffrey, "Approximate Methods in Calculating Discriminant Functions," *Psychometrika,* Vol. 10 (1945), pp. 205–217.
3. Bose, R. C., and Roy, S. N., "The Exact Distribution of the Studentized D^2-Statistic," *Sankhya,* Vol. 4 (1938), pp. 19–31.
4. Cochran, W. G., "The Omission or Addition of an Independent Variate in Multiple Linear Regression," *Supplement to Journal of the Royal Statistical Society,* Vol. V (1938), pp. 171–176.
5. ———., and Bliss, C. I., "Discriminant Functions with Covariance," *Annals of Mathematical Statistics,* Vol. XIX (1948), pp. 151–176.

6. Cox, Gertrude M., and Martin, W., "Use of a Discriminant Function for Differentiating Soils with Different Azotobacter Populations," *Iowa State College Journal of Science*, Vol. XI (1937), pp. 323–331.
7. Day, Besse B., and Sandomire, Marion M., "Use of the Discriminant Function for More Than Two Groups," *Journal of the American Statistical Association*, Vol. 37 (1942), pp. 461–472.
8. Dwyer, P. S., "The Solution of Simultaneous Equations," *Psychometrika*, Vol. 6 (1941), pp. 101–129.
9. ———, "The Implicit Evaluation of Linear Forms and the Solution of Simple Matrix Equations," *Psychometrika*, Vol. 6 (1941), pp. 355–365.
10. Fisher, R. A., "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, Vol. VII (1936), pp. 179–188.
11. ———, "The Statistical Utilization of Multiple Measurements," *Annals of Eugenics*, Vol. VIII (1938), pp. 376–386.
12. ———, "The Precision of Discriminant Functions," *Annals of Eugenics*, Vol. X (1940), pp. 422–429.
13. ———, *Statistical Methods for Research Workers*, 10th ed. Edinburgh: Oliver & Boyd, Ltd., 1946.
14. Horst, Paul, *et al.*, "The Prediction of Personal Adjustment," *Social Science Research Council Bulletin* 48 (1941).
15. Hotelling, Harold, "The Generalization of 'Student's' Ratio," *Annals of Mathematical Statistics*, Vol. II (1931), pp. 360–378.
16. ———, "Relation between Two Sets of Variates," *Biometrika*, Vol. XXVIII (1936), pp. 321–377.
17. ———, "Some New Methods in Matrix Calculation," *Annals of Mathematical Statistics*, Vol. XIV (1943), pp. 1–34.
18. Jackson, Robert W. B., "Approximate Multiple Regression Weights," *Journal of Experimental Education*, Vol. 11 (1943), pp. 221–225.
19. Mahalanobis, P. C., "On the Generalized Distance in Statistics," *Proceedings of the National Institute for Science and Industry*, Vol. 12 (1936), pp. 49–55.
20. Maung, Khint, "Discriminant Analysis of Tocher's Eye-Colour Data for Scottish School Children," *Annals of Eugenics*, Vol. XI (1941), pp. 64–76.
21. Noble, Sister Mary Alfred, Factorial Differentiation by Maximal Difference, *Studies in Psychology and Psychiatry*, Vol. IV (1940). Catholic University Press.
22. Rao, C. Rad., "The Problem of Classification and Distance between Two Populations," *Nature*, Vol. 159 (1947), pp. 30–31.
23. ———, "Tests with Discriminant Functions in Multivariate Analysis," *Sankhya*, Vol. 7 (1946), pp. 407–414.
24. Travers, R. M. W., "The Use of the Discriminant Function in the Treatment of Psychological Group Differences," *Psychometrika*, Vol. 4 (1939), pp. 25–32.
25. Wallace, H. A., and Snedecor, George W., Correlation and Machine Calculation, Iowa State College Press, Vol. XXX (1931).
26. Wald, Abraham, "On a Statistical Problem Arising in the Classification of an Individual into One of Two Groups," *Annals of Mathematical Statistics*, Vol. XV (1944), pp. 145–162.
27. Wherry, Robert J., "Test Selection and Suppressor Variables," *Psychometrika*, Vol. 11 (1946), pp. 239–247.

APPENDIX

TABLE I*

Proportion of the Cases in a Normal Distribution Lying below Certain Values of the Abscissa

Abscissa $\frac{X - M}{S} = z$	Proportion of cases below z	Abscissa $\frac{X - M}{S} = z$	Proportion of cases below z	Abscissa $\frac{X - M}{S} = z$	Proportion of cases below z
.00	.5000	1.25	.8944	2.50	.9938
.05	.5199	1.30	.9032	2.55	.9946
.10	.5398	1.35	.9115	2.60	.9953
.15	.5596	1.40	.9192	2.65	.9960
.20	.5793	1.45	.9265	2.70	.9965
.25	.5987	1.50	.9332	2.75	.9970
.30	.6179	1.55	.9394	2.80	.9974
.35	.6368	1.60	.9452	2.85	.9978
.40	.6554	1.65	.9505	2.90	.9981
.45	.6736	1.70	.9554	2.95	.9984
.50	.6915	1.75	.9599	3.00	.9987
.55	.7088	1.80	.9641	3.05	.9989
.60	.7257	1.85	.9678	3.10	.9990
.65	.7422	1.90	.9713	3.15	.9992
.70	.7580	1.95	.9744	3.20	.9993
.75	.7734	2.00	.9772	3.25	.9994
.80	.7881	2.05	.9798	3.30	.9995
.85	.8023	2.10	.9821	3.35	.9996
.90	.8159	2.15	.9842	3.40	.9997
.95	.8289	2.20	.9861	3.45	.9997
1.00	.8413	2.25	.9878	3.50	.9998
1.05	.8531	2.30	.9893	3.55	.9998
1.10	.8643	2.35	.9906	3.60	.9998
1.15	.8749	2.40	.9918	3.65	.9999
1.20	.8849	2.45	.9929	3.70	.9999

* Table arranged by Dr. Robert W. B. Jackson and used with his permission. For the extended table and for other tables of the normal curve the reader is referred to the tables given by Karl Pearson in *Tables for Statisticians and Biometricians*, Part I, issued by the Biometric Laboratory, University College, London.

TABLE II*

Distribution of *t*

	Probability												
n	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01	.001
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055	3.318
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	.126	.255	.388	.529	.681	.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	.126	.254	.387	.527	.679	.848	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	.126	.254	.386	.526	.677	.845	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	.126	.253	.385	.524	.674	.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

* Table II is reprinted from Table III, Distribution of *t*, in Fisher and Yates, *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

TABLE III*

Distribution of χ^2

Probability

n	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.000157	.000628	.00393	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	.115	.185	.352	.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.268
4	.297	.429	.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.465
5	.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.517
6	.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.001	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.000	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315
21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n - 1}$ may be used as a normal deviate with unit variance.

* Table III is reprinted from Table IV, Distribution of χ^2, in Fisher and Yates, *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the authors and publishers.

TABLE IV*

5% (Roman Type) and 1% (Bold Face Type) Points for the Distribution of *F*

n_2	n_1 degrees of freedom (for greater mean square)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
1	161	200	216	225	230	234	237	239	241	242	243	244	245	246	248	249	250	251	252	253	253	254	254	254	1
	4,052	**4,999**	**5,403**	**5,625**	**5,764**	**5,859**	**5,928**	**5,981**	**6,022**	**6,056**	**6,082**	**6,106**	**6,142**	**6,169**	**6,208**	**6,234**	**6,258**	**6,286**	**6,302**	**6,323**	**6,334**	**6,352**	**6,361**	**6,366**	
2	18.51	19.00	19.16	19.25	19.30	19.33	19.36	19.37	19.38	19.39	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.47	19.47	19.48	19.49	19.49	19.50	19.50	2
	98.49	**99.00**	**99.17**	**99.45**	**99.30**	**99.33**	**99.34**	**99.36**	**99.38**	**99.40**	**99.41**	**99.42**	**99.43**	**99.44**	**99.45**	**99.46**	**99.47**	**99.48**	**99.48**	**99.49**	**99.49**	**99.49**	**99.50**	**99.50**	
3	10.13	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74	8.71	8.69	8.66	8.64	8.62	8.60	8.58	8.57	8.56	8.54	8.54	8.53	3
	34.12	**30.82**	**29.46**	**28.71**	**28.24**	**27.91**	**27.67**	**27.49**	**27.34**	**27.23**	**27.13**	**27.05**	**26.92**	**26.83**	**26.69**	**26.60**	**26.50**	**26.41**	**26.35**	**26.27**	**26.23**	**26.18**	**26.14**	**26.12**	
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.93	5.91	5.87	5.84	5.80	5.77	5.74	5.71	5.70	5.68	5.66	5.65	5.64	5.63	4
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.98**	**14.80**	**14.66**	**14.54**	**14.45**	**14.37**	**14.24**	**14.15**	**14.02**	**13.93**	**13.83**	**13.74**	**13.69**	**13.61**	**13.57**	**13.52**	**13.48**	**13.46**	
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.78	4.74	4.70	4.68	4.64	4.60	4.56	4.53	4.50	4.46	4.44	4.42	4.40	4.38	4.37	4.36	5
	16.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.45**	**10.27**	**10.15**	**10.05**	**9.96**	**9.89**	**9.77**	**9.68**	**9.55**	**9.47**	**9.38**	**9.29**	**9.24**	**9.17**	**9.13**	**9.07**	**9.04**	**9.02**	
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.96	3.92	3.87	3.84	3.81	3.77	3.75	3.72	3.71	3.69	3.68	3.67	6
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.26**	**8.10**	**7.98**	**7.87**	**7.79**	**7.72**	**7.60**	**7.52**	**7.39**	**7.31**	**7.23**	**7.14**	**7.09**	**7.02**	**6.99**	**6.94**	**6.90**	**6.88**	
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57	3.52	3.49	3.44	3.41	3.38	3.34	3.32	3.29	3.28	3.25	3.24	3.23	7
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**7.00**	**6.84**	**6.71**	**6.62**	**6.54**	**6.47**	**6.35**	**6.27**	**6.15**	**6.07**	**5.98**	**5.90**	**5.85**	**5.78**	**5.75**	**5.70**	**5.67**	**5.65**	
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.34	3.31	3.28	3.23	3.20	3.15	3.12	3.08	3.05	3.03	3.00	2.98	2.96	2.94	2.93	8
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.19**	**6.03**	**5.91**	**5.82**	**5.74**	**5.67**	**5.56**	**5.48**	**5.36**	**5.28**	**5.20**	**5.11**	**5.06**	**5.00**	**4.96**	**4.91**	**4.88**	**4.86**	
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.13	3.10	3.07	3.02	2.98	2.93	2.90	2.86	2.82	2.80	2.77	2.76	2.73	2.72	2.71	9
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.62**	**5.47**	**5.35**	**5.26**	**5.18**	**5.11**	**5.00**	**4.92**	**4.80**	**4.73**	**4.64**	**4.56**	**4.51**	**4.45**	**4.41**	**4.36**	**4.33**	**4.31**	
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97	2.94	2.91	2.86	2.82	2.77	2.74	2.70	2.67	2.64	2.61	2.59	2.56	2.55	2.54	10
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.21**	**5.06**	**4.95**	**4.85**	**4.78**	**4.71**	**4.60**	**4.52**	**4.41**	**4.33**	**4.25**	**4.17**	**4.12**	**4.05**	**4.01**	**3.96**	**3.93**	**3.91**	
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.86	2.82	2.79	2.74	2.70	2.65	2.61	2.57	2.53	2.50	2.47	2.45	2.42	2.41	2.40	11
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.88**	**4.74**	**4.63**	**4.54**	**4.46**	**4.40**	**4.29**	**4.21**	**4.10**	**4.02**	**3.94**	**3.86**	**3.80**	**3.74**	**3.70**	**3.66**	**3.62**	**3.60**	
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76	2.72	2.69	2.64	2.60	2.54	2.50	2.46	2.42	2.40	2.36	2.35	2.32	2.31	2.30	12
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.65**	**4.50**	**4.39**	**4.30**	**4.22**	**4.16**	**4.05**	**3.98**	**3.86**	**3.78**	**3.70**	**3.61**	**3.56**	**3.49**	**3.46**	**3.41**	**3.38**	**3.36**	
13	4.67	3.80	3.41	3.18	3.02	2.92	2.84	2.77	2.72	2.67	2.63	2.60	2.55	2.51	2.46	2.42	2.38	2.34	2.32	2.28	2.26	2.24	2.22	2.21	13
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.44**	**4.30**	**4.19**	**4.10**	**4.02**	**3.96**	**3.85**	**3.78**	**3.67**	**3.59**	**3.51**	**3.42**	**3.37**	**3.30**	**3.27**	**3.21**	**3.18**	**3.16**	

n_2	n_1 degrees of freedom (for greater mean square)																							n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60	2.56	2.53	2.48	2.44	2.39	2.35	2.31	2.27	2.24	2.21	2.19	2.16	2.14	2.13	14
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.28**	**4.14**	**4.03**	**3.94**	**3.86**	**3.80**	**3.70**	**3.62**	**3.51**	**3.43**	**3.34**	**3.26**	**3.21**	**3.14**	**3.11**	**3.06**	**3.02**	**3.00**	
15	4.54	3.68	3.29	3.06	2.90	2.79	2.70	2.64	2.59	2.55	2.51	2.48	2.43	2.39	2.33	2.29	2.25	2.21	2.18	2.15	2.12	2.10	2.08	2.07	15
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.14**	**4.00**	**3.89**	**3.80**	**3.73**	**3.67**	**3.56**	**3.48**	**3.36**	**3.29**	**3.20**	**3.12**	**3.07**	**3.00**	**2.97**	**2.92**	**2.89**	**2.87**	
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.45	2.42	2.37	2.33	2.28	2.24	2.20	2.16	2.13	2.09	2.07	2.04	2.02	2.01	16
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**4.03**	**3.89**	**3.78**	**3.69**	**3.61**	**3.55**	**3.45**	**3.37**	**3.25**	**3.18**	**3.10**	**3.01**	**2.96**	**2.89**	**2.86**	**2.80**	**2.77**	**2.75**	
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.50	2.45	2.41	2.38	2.33	2.29	2.23	2.19	2.15	2.11	2.08	2.04	2.02	1.99	1.97	1.96	17
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.93**	**3.79**	**3.68**	**3.59**	**3.52**	**3.45**	**3.35**	**3.27**	**3.16**	**3.08**	**3.00**	**2.92**	**2.86**	**2.79**	**2.76**	**2.70**	**2.67**	**2.65**	
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.29	2.25	2.19	2.15	2.11	2.07	2.04	2.00	1.98	1.95	1.93	1.92	18
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.85**	**3.71**	**3.60**	**3.51**	**3.44**	**3.37**	**3.27**	**3.19**	**3.07**	**3.00**	**2.91**	**2.83**	**2.78**	**2.71**	**2.68**	**2.62**	**2.59**	**2.57**	
19	4.38	3.52	3.13	2.90	2.74	2.63	2.55	2.48	2.43	2.38	2.34	2.31	2.26	2.21	2.15	2.11	2.07	2.02	2.00	1.96	1.94	1.91	1.90	1.88	19
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.77**	**3.63**	**3.52**	**3.43**	**3.36**	**3.30**	**3.19**	**3.12**	**3.00**	**2.92**	**2.84**	**2.76**	**2.70**	**2.63**	**2.60**	**2.54**	**2.51**	**2.49**	
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35	2.31	2.28	2.23	2.18	2.12	2.08	2.04	1.99	1.96	1.92	1.90	1.87	1.85	1.84	20
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.71**	**3.56**	**3.45**	**3.37**	**3.30**	**3.23**	**3.13**	**3.05**	**2.94**	**2.86**	**2.77**	**2.69**	**2.63**	**2.56**	**2.53**	**2.47**	**2.44**	**2.42**	
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25	2.20	2.15	2.09	2.05	2.00	1.96	1.93	1.89	1.87	1.84	1.82	1.81	21
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.65**	**3.51**	**3.40**	**3.31**	**3.24**	**3.17**	**3.07**	**2.99**	**2.88**	**2.80**	**2.72**	**2.63**	**2.58**	**2.51**	**2.47**	**2.42**	**2.38**	**2.36**	
22	4.30	3.44	3.05	2.82	2.66	2.55	2.47	2.40	2.35	2.30	2.26	2.23	2.18	2.13	2.07	2.03	1.98	1.93	1.91	1.87	1.84	1.81	1.80	1.78	22
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.76**	**3.59**	**3.45**	**3.35**	**3.26**	**3.18**	**3.12**	**3.02**	**2.94**	**2.83**	**2.75**	**2.67**	**2.58**	**2.53**	**2.46**	**2.42**	**2.37**	**2.33**	**2.31**	
23	4.28	3.42	3.03	2.80	2.64	2.53	2.45	2.38	2.32	2.28	2.24	2.20	2.14	2.10	2.04	2.00	1.96	1.91	1.88	1.84	1.82	1.79	1.77	1.76	23
	7.88	**5.66**	**4.76**	**4.26**	**3.94**	**3.71**	**3.54**	**3.41**	**3.30**	**3.21**	**3.14**	**3.07**	**2.97**	**2.89**	**2.78**	**2.70**	**2.62**	**2.53**	**2.48**	**2.41**	**2.37**	**2.32**	**2.28**	**2.26**	
24	4.26	3.40	3.01	2.78	2.62	2.51	2.43	2.36	2.30	2.26	2.22	2.18	2.13	2.09	2.02	1.98	1.94	1.89	1.86	1.82	1.80	1.76	1.74	1.73	24
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.50**	**3.36**	**3.25**	**3.17**	**3.09**	**3.03**	**2.93**	**2.85**	**2.74**	**2.66**	**2.58**	**2.49**	**2.44**	**2.36**	**2.33**	**2.27**	**2.23**	**2.21**	
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24	2.20	2.16	2.11	2.06	2.00	1.96	1.92	1.87	1.84	1.80	1.77	1.74	1.72	1.71	25
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.46**	**3.32**	**3.21**	**3.13**	**3.05**	**2.99**	**2.89**	**2.81**	**2.70**	**2.62**	**2.54**	**2.45**	**2.40**	**2.32**	**2.29**	**2.23**	**2.19**	**2.17**	
26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.10	2.05	1.99	1.95	1.90	1.85	1.82	1.78	1.76	1.72	1.70	1.69	26
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.42**	**3.29**	**3.17**	**3.09**	**3.02**	**2.96**	**2.86**	**2.77**	**2.66**	**2.58**	**2.50**	**2.41**	**2.36**	**2.28**	**2.25**	**2.19**	**2.15**	**2.13**	

The function, $F = e$ with exponent $2z$, is computed in part from Fisher's table VI (7). Additional entries are by interpolation, mostly graphical.

* This table is reproduced from Table 10.7 in *Statistical Methods* (4th ed.), 1946, with the permission of Professor George W. Snedecor and the publishers, The Iowa State College Press, Ames, Iowa.

TABLE IV* (*Continued*)

n_2	n_1 degrees of freedom (for greater mean square)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	2.20	2.16	2.13	2.08	2.03	1.97	1.93	1.88	1.84	1.80	1.76	1.74	1.71	1.68	1.67	27
	7.68	**5.49**	**4.60**	**4.11**	**3.79**	**3.56**	**3.39**	**3.26**	**3.14**	**3.06**	**2.98**	**2.93**	**2.83**	**2.74**	**2.63**	**2.55**	**2.47**	**2.38**	**2.33**	**2.25**	**2.21**	**2.16**	**2.12**	**2.10**	
28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	2.19	2.15	2.12	2.06	2.02	1.96	1.91	1.87	1.81	1.78	1.75	1.72	1.69	1.67	1.65	28
	7.64	**5.45**	**4.57**	**4.07**	**3.76**	**3.53**	**3.36**	**3.23**	**3.11**	**3.03**	**2.95**	**2.90**	**2.80**	**2.71**	**2.60**	**2.52**	**2.44**	**2.35**	**2.30**	**2.22**	**2.18**	**2.13**	**2.09**	**2.06**	
29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	2.18	2.14	2.10	2.05	2.00	1.94	1.90	1.85	1.80	1.77	1.73	1.71	1.68	1.65	1.64	29
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.33**	**3.20**	**3.08**	**3.00**	**2.92**	**2.87**	**2.77**	**2.68**	**2.57**	**2.49**	**2.41**	**2.32**	**2.27**	**2.19**	**2.15**	**2.10**	**2.06**	**2.03**	
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16	2.12	2.09	2.04	1.99	1.93	1.89	1.84	1.79	1.76	1.72	1.69	1.66	1.64	1.62	30
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.30**	**3.17**	**3.06**	**2.98**	**2.90**	**2.84**	**2.74**	**2.66**	**2.55**	**2.47**	**2.38**	**2.29**	**2.24**	**2.16**	**2.13**	**2.07**	**2.03**	**2.01**	
32	4.15	3.30	2.90	2.67	2.51	2.40	2.32	2.25	2.19	2.14	2.10	2.07	2.02	1.97	1.91	1.86	1.82	1.76	1.74	1.69	1.67	1.64	1.61	1.59	32
	7.50	**5.34**	**4.46**	**3.97**	**3.66**	**3.42**	**3.25**	**3.12**	**3.01**	**2.94**	**2.86**	**2.80**	**2.70**	**2.62**	**2.51**	**2.42**	**2.34**	**2.25**	**2.20**	**2.12**	**2.08**	**2.02**	**1.98**	**1.96**	
34	4.13	3.28	2.88	2.65	2.49	2.38	2.30	2.23	2.17	2.12	2.08	2.05	2.00	1.95	1.89	1.84	1.80	1.74	1.71	1.67	1.64	1.61	1.59	1.57	34
	7.44	**5.29**	**4.42**	**3.93**	**3.61**	**3.38**	**3.21**	**3.08**	**2.97**	**2.89**	**2.82**	**2.76**	**2.66**	**2.58**	**2.47**	**2.38**	**2.30**	**2.21**	**2.15**	**2.08**	**2.04**	**1.98**	**1.94**	**1.91**	
36	4.11	3.26	2.86	2.63	2.48	2.36	2.28	2.21	2.15	2.10	2.06	2.03	1.98	1.93	1.87	1.82	1.78	1.72	1.69	1.65	1.62	1.59	1.56	1.55	36
	7.39	**5.25**	**4.38**	**3.89**	**3.58**	**3.35**	**3.18**	**3.04**	**2.94**	**2.86**	**2.78**	**2.72**	**2.62**	**2.54**	**2.43**	**2.35**	**2.26**	**2.17**	**2.12**	**2.04**	**2.00**	**1.94**	**1.90**	**1.87**	
38	4.10	3.25	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02	1.96	1.92	1.85	1.80	1.76	1.71	1.67	1.63	1.60	1.57	1.54	1.53	38
	7.35	**5.21**	**4.34**	**3.86**	**3.54**	**3.32**	**3.15**	**3.02**	**2.91**	**2.82**	**2.75**	**2.69**	**2.59**	**2.51**	**2.40**	**2.32**	**2.22**	**2.14**	**2.08**	**2.00**	**1.97**	**1.90**	**1.86**	**1.84**	
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07	2.04	2.00	1.95	1.90	1.84	1.79	1.74	1.69	1.66	1.61	1.59	1.55	1.53	1.51	40
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**3.12**	**2.99**	**2.88**	**2.80**	**2.73**	**2.66**	**2.56**	**2.49**	**2.37**	**2.29**	**2.20**	**2.11**	**2.05**	**1.97**	**1.94**	**1.88**	**1.84**	**1.81**	
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.02	1.99	1.94	1.89	1.82	1.78	1.73	1.68	1.64	1.60	1.57	1.54	1.51	1.49	42
	7.27	**5.15**	**4.29**	**3.80**	**3.49**	**3.26**	**3.10**	**2.96**	**2.86**	**2.77**	**2.70**	**2.64**	**2.54**	**2.46**	**2.35**	**2.26**	**2.17**	**2.08**	**2.02**	**1.94**	**1.91**	**1.85**	**1.80**	**1.78**	
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98	1.92	1.88	1.81	1.76	1.72	1.66	1.63	1.58	1.56	1.52	1.50	1.48	44
	7.24	**5.12**	**4.26**	**3.78**	**3.46**	**3.24**	**3.07**	**2.94**	**2.84**	**2.75**	**2.68**	**2.62**	**2.52**	**2.44**	**2.32**	**2.24**	**2.15**	**2.06**	**2.00**	**1.92**	**1.88**	**1.82**	**1.78**	**1.75**	
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.14	2.09	2.04	2.00	1.97	1.91	1.87	1.80	1.75	1.71	1.65	1.62	1.57	1.54	1.51	1.48	1.46	46
	7.21	**5.10**	**4.24**	**3.76**	**3.44**	**3.22**	**3.05**	**2.92**	**2.82**	**2.73**	**2.66**	**2.60**	**2.50**	**2.42**	**2.30**	**2.22**	**2.13**	**2.04**	**1.98**	**1.90**	**1.86**	**1.80**	**1.76**	**1.72**	
48	4.04	3.19	2.80	2.56	2.41	2.30	2.21	2.14	2.08	2.03	1.99	1.96	1.90	1.86	1.79	1.74	1.70	1.64	1.61	1.56	1.53	1.50	1.47	1.45	48
	7.19	**5.08**	**4.22**	**3.74**	**3.42**	**3.20**	**3.04**	**2.90**	**2.80**	**2.71**	**2.64**	**2.58**	**2.48**	**2.40**	**2.28**	**2.20**	**2.11**	**2.02**	**1.96**	**1.88**	**1.84**	**1.78**	**1.73**	**1.70**	

n_2	n_1 degrees of freedom (for greater mean square)																								n_2
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20	24	30	40	50	75	100	200	500	∞	
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02	1.98	1.95	1.90	1.85	1.78	1.74	1.69	1.63	1.60	1.55	1.52	1.48	1.46	1.44	50
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.18**	**3.02**	**2.88**	**2.78**	**2.70**	**2.62**	**2.56**	**2.46**	**2.39**	**2.26**	**2.18**	**2.10**	**2.00**	**1.94**	**1.86**	**1.82**	**1.76**	**1.71**	**1.68**	
55	4.02	3.17	2.78	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.97	1.93	1.88	1.83	1.76	1.72	1.67	1.61	1.58	1.52	1.50	1.46	1.43	1.41	55
	7.12	**5.01**	**4.16**	**3.68**	**3.37**	**3.15**	**2.98**	**2.85**	**2.75**	**2.66**	**2.59**	**2.53**	**2.43**	**2.35**	**2.23**	**2.15**	**2.06**	**1.96**	**1.90**	**1.82**	**1.78**	**1.71**	**1.66**	**1.64**	
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.86	1.81	1.75	1.70	1.65	1.59	1.56	1.50	1.48	1.44	1.41	1.39	60
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.95**	**2.82**	**2.72**	**2.63**	**2.56**	**2.50**	**2.40**	**2.32**	**2.20**	**2.12**	**2.03**	**1.93**	**1.87**	**1.79**	**1.74**	**1.68**	**1.63**	**1.60**	
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.02	1.98	1.94	1.90	1.85	1.80	1.73	1.68	1.63	1.57	1.54	1.49	1.46	1.42	1.39	1.37	65
	7.04	**4.95**	**4.10**	**3.62**	**3.31**	**3.09**	**2.93**	**2.79**	**2.70**	**2.61**	**2.54**	**2.47**	**2.37**	**2.30**	**2.18**	**2.09**	**2.00**	**1.90**	**1.84**	**1.76**	**1.71**	**1.64**	**1.60**	**1.56**	
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.01	1.97	1.93	1.89	1.84	1.79	1.72	1.67	1.62	1.56	1.53	1.47	1.45	1.40	1.37	1.35	70
	7.01	**4.92**	**4.08**	**3.60**	**3.29**	**3.07**	**2.91**	**2.77**	**2.67**	**2.59**	**2.51**	**2.45**	**2.35**	**2.28**	**2.15**	**2.07**	**1.98**	**1.88**	**1.82**	**1.74**	**1.69**	**1.62**	**1.56**	**1.53**	
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95	1.91	1.88	1.82	1.77	1.70	1.65	1.60	1.54	1.51	1.45	1.42	1.38	1.35	1.32	80
	6.96	**4.88**	**4.04**	**3.56**	**3.25**	**3.04**	**2.87**	**2.74**	**2.64**	**2.55**	**2.48**	**2.41**	**2.32**	**2.24**	**2.11**	**2.03**	**1.94**	**1.84**	**1.78**	**1.70**	**1.65**	**1.57**	**1.52**	**1.49**	
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92	1.88	1.85	1.79	1.75	1.68	1.63	1.57	1.51	1.48	1.42	1.39	1.34	1.30	1.28	100
	6.90	**4.82**	**3.98**	**3.51**	**3.20**	**2.99**	**2.82**	**2.69**	**2.59**	**2.51**	**2.43**	**2.36**	**2.26**	**2.19**	**2.06**	**1.98**	**1.89**	**1.79**	**1.73**	**1.64**	**1.59**	**1.51**	**1.46**	**1.43**	
125	3.92	3.07	2.68	2.44	2.29	2.17	2.08	2.01	1.95	1.90	1.86	1.83	1.77	1.72	1.65	1.60	1.55	1.49	1.45	1.39	1.36	1.31	1.27	1.25	125
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.79**	**2.65**	**2.56**	**2.47**	**2.40**	**2.33**	**2.23**	**2.15**	**2.03**	**1.94**	**1.85**	**1.75**	**1.68**	**1.59**	**1.54**	**1.46**	**1.40**	**1.37**	
150	3.91	3.06	2.67	2.43	2.27	2.16	2.07	2.00	1.94	1.89	1.85	1.82	1.76	1.71	1.64	1.59	1.54	1.47	1.44	1.37	1.34	1.29	1.25	1.22	150
	6.81	**4.75**	**3.91**	**3.44**	**3.14**	**2.92**	**2.76**	**2.62**	**2.53**	**2.44**	**2.37**	**2.30**	**2.20**	**2.12**	**2.00**	**1.91**	**1.83**	**1.72**	**1.66**	**1.56**	**1.51**	**1.43**	**1.37**	**1.33**	
200	3.89	3.04	2.65	2.41	2.26	2.14	2.05	1.98	1.92	1.87	1.83	1.80	1.74	1.69	1.62	1.57	1.52	1.45	1.42	1.35	1.32	1.26	1.22	1.19	200
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.90**	**2.73**	**2.60**	**2.50**	**2.41**	**2.34**	**2.28**	**2.17**	**2.09**	**1.97**	**1.88**	**1.79**	**1.69**	**1.62**	**1.53**	**1.48**	**1.39**	**1.33**	**1.28**	
400	3.86	3.02	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.81	1.78	1.72	1.67	1.60	1.54	1.49	1.42	1.38	1.32	1.28	1.22	1.16	1.13	400
	6.70	**4.66**	**3.83**	**3.36**	**3.06**	**2.85**	**2.69**	**2.55**	**2.46**	**2.37**	**2.29**	**2.23**	**2.12**	**2.04**	**1.92**	**1.84**	**1.74**	**1.64**	**1.57**	**1.47**	**1.42**	**1.32**	**1.24**	**1.19**	
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84	1.80	1.76	1.70	1.65	1.58	1.53	1.47	1.41	1.36	1.30	1.26	1.19	1.13	1.08	1000
	6.66	**4.62**	**3.80**	**3.34**	**3.04**	**2.82**	**2.66**	**2.53**	**2.43**	**2.34**	**2.26**	**2.20**	**2.09**	**2.01**	**1.89**	**1.81**	**1.71**	**1.61**	**1.54**	**1.44**	**1.38**	**1.28**	**1.19**	**1.11**	
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83	1.79	1.75	1.69	1.64	1.57	1.52	1.46	1.40	1.35	1.28	1.24	1.17	1.11	1.00	∞
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.64**	**2.51**	**2.41**	**2.32**	**2.24**	**2.18**	**2.07**	**1.99**	**1.87**	**1.79**	**1.69**	**1.59**	**1.52**	**1.41**	**1.36**	**1.25**	**1.15**	**1.00**	

The function, $F = e$ with exponent $2z$, is computed in part from Fisher's table VI (7). Additional entries are by interpolation, mostly graphical.

* This table is reproduced from Table 10.7 in *Statistical Methods* (4th ed.), 1946, with the permission of Professor George W. Snedecor and the publishers, The Iowa State College Press, Ames, Iowa.

TABLE V*

Test for Significance of Differences in Variance among K Samples Each of Size n (P. P. N. Nayer's Tables)

5% limits of L_1

f	1	2	3	4	5	6	7	8	9	11	14	19	29	59	∞
k \ n	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
2	.079	.312	.478	.585	.656	.708	.745	.775	.798	.833	.868	.902	.935	.968	1.000
3		.304	.470	.576	.648	.700	.739	.769	.792	.828	.863	.898	.933	.967	1.000
4		.315	.480	.585	.656	.707	.744	.774	.797	.832	.866	.900	.934	.967	1.000
5		.328	.491	.595	.665	.714	.751	.780	.802	.836	.870	.903	.936	.968	1.000
6		.339	.502	.604	.673	.721	.757	.785	.808	.841	.873	.906	.938	.969	1.000
7		.350	.512	.612	.680	.727	.763	.790	.812	.844	.876	.908	.939	.970	1.000
8		.359	.520	.620	.686	.733	.768	.795	.816	.848	.879	.910	.941	.971	1.000
9		.367	.527	.626	.691	.738	.772	.798	.819	.851	.881	.912	.942	.971	1.000
10	.117	.374	.534	.631	.696	.742	.776	.802	.822	.853	.883	.913	.943	.972	1.000
12	.124	.387	.545	.641	.704	.749	.782	.807	.828	.857	.887	.916	.944	.973	1.000
14	.130	.397	.554	.649	.711	.755	.787	.812	.832	.861	.890	.918	.946	.973	1.000
16	.136	.405	.561	.655	.716	.759	.791	.816	.835	.863	.892	.920	.947	.974	1.000
18	.142	.412	.567	.660	.721	.763	.795	.819	.838	.866	.894	.921	.948	.974	1.000
20	.147	.418	.573	.665	.725	.767	.798	.822	.840	.868	.896	.922	.949	.975	1.000
22	.152	.424	.577	.669	.728	.770	.800	.824	.843	.870	.897	.924	.950	.975	1.000
24	.156	.428	.581	.672	.731	.772	.802	.826	.844	.872	.898	.924	.950	.975	1.000
26	.160	.433	.585	.675	.734	.775	.805	.828	.846	.873	.899	.925	.951	.976	1.000
28	.163	.437	.589	.678	.736	.777	.807	.829	.848	.874	.900	.926	.951	.976	1.000
30	.166	.441	.592	.681	.739	.779	.809	.831	.849	.876	.901	.927	.952	.976	1.000

Criterion $L_1 = \dfrac{\prod_t (S_t^2)^{\frac{1}{k}}}{\frac{1}{k}\sum_t (S_t^2)}$, where $S_t^2 = \frac{1}{r_t}\sum (X - \bar{X}_t)^2$. *Note:* For $n = 2$, $k = 50$ and $L_{.05} = .187$.

TABLE V (*Continued*)

1% limits of L_1

f	1	2	3	4	5	6	7	8	9	11	14	19	29	59	∞
k \ *n*	2	3	4	5	6	7	8	9	10	12	15	20	30	60	∞
2	.016	.141	.284	.398	.485	.551	.603	.645	.678	.730	.783	.836	.890	.945	1.000
3		.162	.314	.429	.514	.578	.628	.667	.699	.748	.798	.848	.898	.949	1.000
4		.188	.345	.459	.542	.604	.652	.689	.719	.765	.812	.859	.906	.953	1.000
5		.210	.370	.484	.565	.624	.670	.706	.735	.779	.823	.867	.911	.956	1.000
6		.230	.391	.504	.583	.641	.685	.720	.748	.789	.832	.874	.916	.958	1.000
7		.246	.409	.520	.597	.654	.697	.730	.757	.798	.839	.879	.920	.960	1.000
8		.260	.424	.534	.610	.665	.707	.740	.766	.805	.844	.884	.923	.962	1.000
9		.273	.437	.545	.620	.674	.715	.747	.773	.811	.849	.887	.925	.963	1.000
10	.063	.284	.448	.555	.629	.682	.722	.753	.779	.816	.853	.890	.927	.964	1.000
12	.071	.303	.467	.572	.644	.696	.734	.764	.789	.824	.860	.896	.931	.966	1.000
14	.079	.318	.481	.585	.655	.706	.744	.773	.796	.831	.865	.900	.933	.967	1.000
16	.087	.331	.493	.596	.665	.714	.751	.779	.802	.836	.870	.903	.936	.968	1.000
18	.093	.342	.504	.605	.672	.721	.756	.784	.807	.840	.873	.905	.937	.969	1.000
20	.100	.352	.512	.613	.679	.727	.761	.788	.811	.844	.876	.908	.939	.970	1.000
22	.105	.360	.520	.619	.684	.732	.765	.792	.814	.847	.878	.909	.940	.970	1.000
24	.110	.367	.526	.624	.688	.736	.768	.795	.817	.850	.880	.911	.941	.971	1.000
26	.115	.373	.532	.629	.693	.740	.772	.798	.820	.852	.882	.912	.942	.971	1.000
28	.119	.379	.537	.634	.697	.744	.776	.802	.823	.854	.884	.914	.943	.972	1.000
30	.123	.386	.543	.639	.703	.748	.781	.806	.827	.856	.886	.915	.944	.972	1.000

Note: For $n = 2$, $k = 50$ and $L_{.01} = .151$.

* These tables are reproduced, by permission, from ***Statistical Research Memoirs***, Vol. I, edited by J. Neyman and E. S. Pearson and issued by the Department of Statistics, University of London, University College, London.

INDEX

INDEX